How far will your favorite basketball team go in NCAA March Madness?

Merge Conflict: Cassidy R., Ella T., Ethan S., Pierre Q.

2025-04-28

Introduction

Arguably the most uniting part of Duke is basketball. With lively traditions, incredible game-day culture, and overall a stellar basketball team, what's not to love?! That being said, one thing that unites Duke students even more is friendly competition. Consequently, on our mission to combine the two, we stumble upon an annual tradition: The NCAA March Madness Tournament… more specifically, the challenge of perfecting a bracket. Using our knowledge from STA221, we believe that we have been equipped with a skill set to more accurately predict how far a team will go in the tournament based on historical data.

The NCAA March Madness tournament offers a wealth of sports analytics data for data scientists to analyze. Individual and team statistics go into advanced metrics that are used to rate and predict the performance of teams in the tournament. Part of tournament is the opportunity to create a bracket that predicts the results of the single elimination tournament, but the history of the bracket challenge, there has never been a perfect bracket. If we were to flip a coin to pick the winner, we would have 1 in 2^{63} or 1 in 9,223,372,036,854,775,808 odds. If we improve our strategy and follow historical winning odds (e.g. how a 1 seed and a 16 seed have performed against each other), we would improve to 1 in 46,576,549,017 - a lot better, but that's still astronomical - about the same amount of days left until the star explodes. Several studies have explored methods to improve March Madness predictions using advanced analytics. For instance, Singh et al. (2023) proposed a deep learning-based approach in their study Advancing NCAA March Madness Forecasts Through Deep Learning and Combinatorial Fusion Analysis, but we will not be utilizing a machine learning approach, instead focusing on models based on historical advanced metrics.

Given the absurd difficulty of getting a perfect bracket, we can ask the question: How accurately can a predictive model based on historical performance metrics, such as offensive and defensive ratings from Ken Pomeroy's analytics, forecast NCAA March Madness outcomes compared to traditional seeding and random selection methods?

Dataset

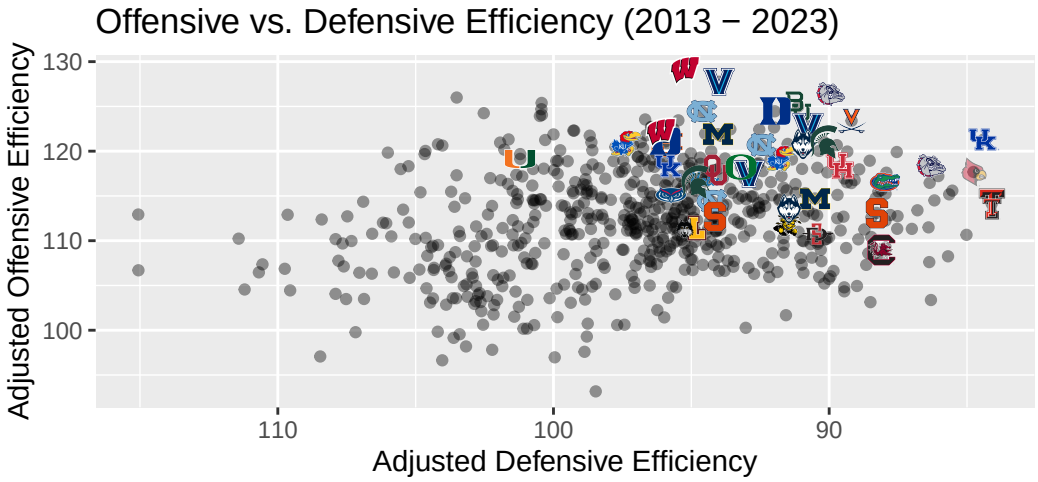
We sourced our data from Kaggle, using a dataset authored by Nishaan Amin, which holds a large collection of different statistics, rankings, and ratings from multiple sites. For our analysis we will be using statistics from kenpom.com using Pomeroy College Basketball advanced metrics dating back to 2008 and yearly tournament index statistics since 2013 from heatcheckcbb.com.

Variable	Name	Description
`ROUND`	Round Reached	Furthest round a team advances to in the tournament.
`AdjOE`	Adjusted Offensive Efficiency	Points scored per 100 possessions, adjusted for opponent strength.
`AdjDE`	Adjusted Defensive Efficiency	Points allowed per 100 possessions, adjusted for opponent strength.
`AdjEM`	Adjusted Efficiency Margin	Expected point margin against the average team (`AdjOE` - `AdjDE`).
`AdjTempo`	Adjusted Tempo	Number of possessions per game (per 40 minutes), adjusted for opponent.
`Off.eFGPct`	Effective Field Goal Percentage	Shooting efficiency (accounting for relative 3-point value).
`Def.eFGPct`	Effective Field Goal Percentage (Opponent)	Shooting efficiency of opponent (accounting for relative 3-point value)
`BlockPct`	Block Percentage	Percentage of opponent field goal attempts blocked.
`StlRate`	Steal Rate	Rate of opponent possessions ending in a steal.
`Experience`	Experience	Average number of years of experience among players.
`AvgHeight`	Average Height	Average height (in inches) among players.
`Off2PtFG`	2pt Field Goals Made	Average number of 2-point shots made a game.
`Off3PtFG`	3pt Field Goals Made	Average number of 3-point shots made a game.

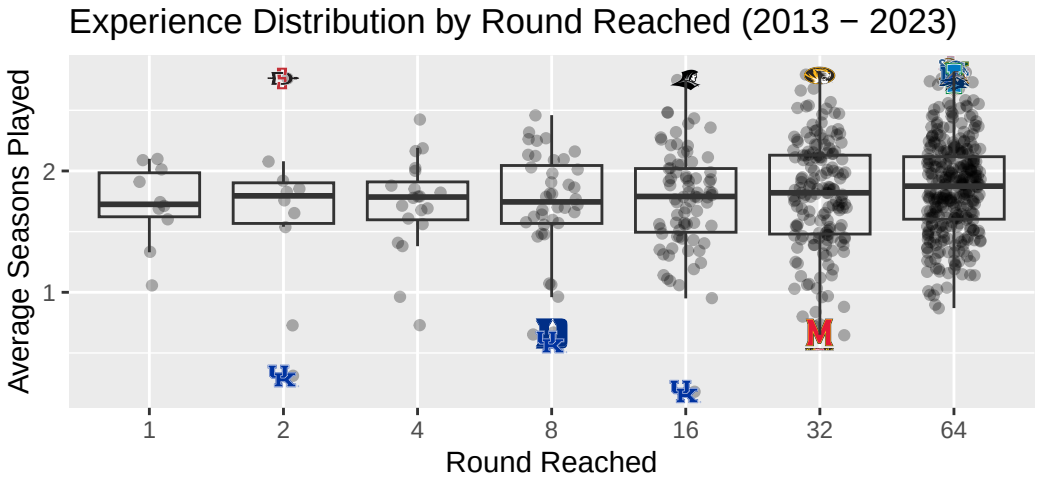
Our analysis will attempt to predict `ROUND`, the furthest round a team reaches in the tournament. It's a factor with levels 64, 32, 16, 8, 4, 2, 1, where 64 represents a loss in the first round, and 1 represents a championship win.

Exploratory Data Analysis

Let's take a look at initial visualizations on how predictors can impact our response, `ROUND`.



This plot takes a look at offensive and defensive efficiency (the former increases for great offensive teams, the second decreases for great defensive teams), with logos indicating a final four appearance. This visualization already indicates that these two may be solid predictors, as most final four teams are good defensive and offensive teams based on these two metrics.



Next we can look at the distribution of team experience by round reached, outliers shown with team logos. Successful teams seem to have a balanced amount of experience, meaning that this may be a weaker predictor, with notable exceptions like the freshman talent squads of Kentucky and Duke and the veteran-based SDSU and Providence teams.

Methodology

Linear

We began our modeling stage with an initial linear model with all terms and will try to pare it down using different model metrics, while making sure our model retains predictive ability. We'll begin by checking for multicollinearity among predictors.

	AdjOE	AdjDE	AdjTempo	Off.eFGPct	Def.eFGPct
withEM	3.153803e+08	2.623516e+08	1.141	2.521	3.074
without	2.237000e+00	3.383000e+00	1.140	2.518	3.070

From initial examination, `AdjOE`, `AdjDE`, and `AdjEM` are extremely correlated with values on the order of 10^8, which makes sense since `AdjEM` is the difference of the others. We'll address this by removing AdjEM in our model, which solves our multicollinearity issue (all predictor VIFs fall below our threshold for concern of 5). Next, we can identify weak predictors by inference using p-values. Most were not statistically significant, as shown in the following table:

term	estimate	std.error	statistic	p.value
Off.eFGPct	0.019	0.415	0.046	0.963
AdjTempo	0.015	0.247	0.061	0.952
Def.eFGPct	-0.110	0.530	-0.208	0.835
AvgHeight	0.367	1.039	0.353	0.724
Off3PtFG	-0.123	0.322	-0.382	0.702
BlockPct	0.209	0.336	0.622	0.534
Off2PtFG	-0.289	0.353	-0.818	0.413
Experience	1.645	1.794	0.917	0.360
(Intercept)	90.859	89.637	1.014	0.311
StlRate	-75.082	49.101	-1.529	0.127

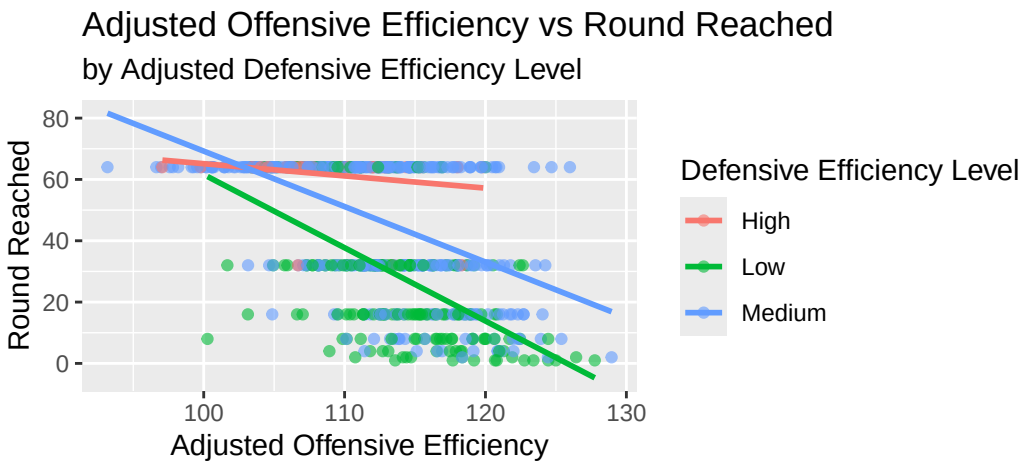
Therefore, we tried a significantly reduced model, leaving `AdjOE` and `AdjDE` as our only 2 predictors. With an AIC 11.76 lower and a BIC 11.76 lower than the previous model, we felt that the removal was justified, but also conducted a drop-in deviance test to confirm.

term	df.residual	df	sumsq	p.value
ROUND ~ AdjOE + AdjDE	629	NA	NA	NA
ROUND ~ AdjOE + AdjDE + Adj…	620	9	1936.718	0.725

With our removed variables not being statistically significant, we can finalize our model to just `AdjOE` and `AdjDE`.

Interaction

As we analyzed the impact that certain factors have on round-reached through modeling, we wanted to explore potential interaction between variables and how that might improve our model. We took a look at interaction terms between Adjusted Offensive and Adjusted Defensive Efficiency since we were interested in how the combination of a team's scoring ability and defensive strength might jointly influence tournament success, especially as we observed that they were the most important predictors for our linear model.



To visualize, we created levels to observe how low, medium, and high Adjusted Defensive Efficiency affects the relationship between Adjusted Offensive Efficiency and round reached. For teams with low defensive efficiency, offensive efficiency plays a much stronger role in determining how far they advance - we can see that this has the steepest line and highest slope on the graph. Teams with high or medium defensive strength appear less dependent on offensive ratings. The non-parallel lines highlight a meaningful interaction between adjusted offensive efficiency and adjusted defensive efficiency. To make sure, we can use a drop-in deviance test to confirm the utility of this interaction.

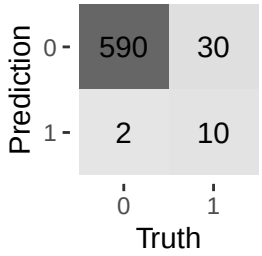
term	df.residual	rss	df	sumsq	p.value
ROUND ~ AdjOE + AdjDE	629	197220.5	NA	NA	NA
ROUND ~ AdjOE * AdjDE	628	186046.5	1	11174.02	0

With a statistically significant p-value, we can reject the null hypothesis that the interaction term is 0, and confirm the choice of adding it.

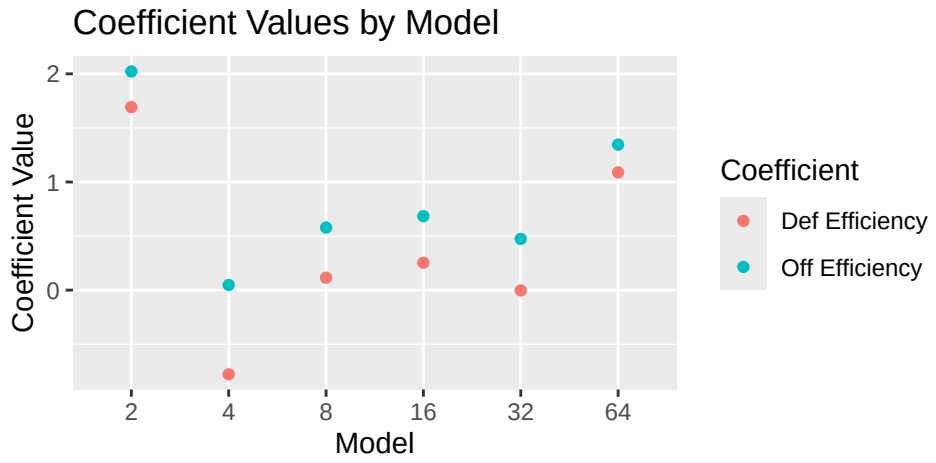
We also checked assumptions with this complete model (see Appendix), which unfortunately were not entirely verified, likely because `ROUND` is a factor instead of numerical. We can still derive a lot of value out of this model, but this pushed us towards a logistic one as well.

Logistic

Logistic models are meant to predict binary results, which seems ill-suited for our task, where we're attempting to figure out how far a team goes in the tournament. A first thought is to try to build a model predicting if a team will make it to the Final Four, for instance. But if we attempt to, we come across the essential problem of data skew - out of 64 (give or take) teams in each tournament, only 4 make it to semi-final, meaning that our initial logistic model is hugely prone to data skew, with an accuracy of $\frac{600}{632} \approx 95\%$, but a sensitivity of $\frac{10}{40} = 25\%$.



So, is there any way to reduce this data disparity? Well, one important feature of the tournament is that rounds are sequential, such that you have to win a game in the round of 32 to proceed to the sweet 16, for instance. So what if we make a model for each round, trimming the data down as we go through to those who made it that far? For our "made it to the sweet sixteen" model, we could then only consider the 32 teams that one the first game. This solves our data skew, with an even 50/50 split of winners and losers each round. It also addresses a fact we didn't initially think about: each round is fundamentally different, with distinct match-ups, locations, and other factors that can change in whose favor a game swings.



Indeed, the above plot shows how the values of our logistic regression models (which use the interaction between offensive and defensive efficiency) change by round, confirming this idea of the difference between rounds. Interestingly, it seems that offense is most important in the championship game and in the round of 64, whereas defense matters more in the intermediary rounds.

For predictions, we'll iterate through each of these 6 models, stopping when the team loses a game. Initially, we'll start with thresholds of 0.5 for each model for an initial check to see if this method is even viable.

```
[1] "Virginia (2019) Actual Result: 1"
[1] "------------- ROUND BY ROUND RESULTS -------------"
[1] "64 -> 32 : 99.34 % TRUE"
[1] "32 -> 16 : 95.47 % TRUE"
[1] "16 -> 8 : 87.09 % TRUE"
[1] "8 -> 4 : 70.14 % TRUE"
[1] "4 -> 2 : 61.42 % TRUE"
[1] "2 -> 1 : 51.3 % TRUE"
[1] "-------------- END OF TOURNAMENT --------------"
[1] "Result: Reached Round of 1"
```

According to our initial results, it does seem to work! Thresholds will be adjusted according to ROC curves, but teams that went high in the tournament (such as 2019 Virginia) are already being predicted to succeed.

We also checked assumptions with these new logistic models (see Appendix), and they were absolutely verified, with a linear relationship between empirical logit and our predictors. Independence and randomness were a little trickier because teams are ranked based on performance against each other, but we considered our solid sample size to be enough to account for this.

Results

Linear

Our final linear model is one that uses `AdjOE`, `AdjDE`, as well as their interaction to predict the round reached of a March Madness team. All predictors are statistically significant with p-values of approximately zero.

term	estimate	std.error	statistic	p.value
(Intercept)	1691.179	261.959	6.456	0
AdjOE	-16.144	2.347	-6.880	0
AdjDE	-14.836	2.674	-5.548	0
AdjOE:AdjDE	0.147	0.024	6.141	0

This model has extremely low multicollinearity among variables with VIF's near 1 for both `AdjOE` and `AdjDE`, as well as the lowest AIC and BIC values of all the models that we tested (5396.37 and 5418.62). The linear model with the predictors `AdjOE`, `AdjDE`, and their interaction is therefore the best fit for predicting the round reached of a team in March Madness.

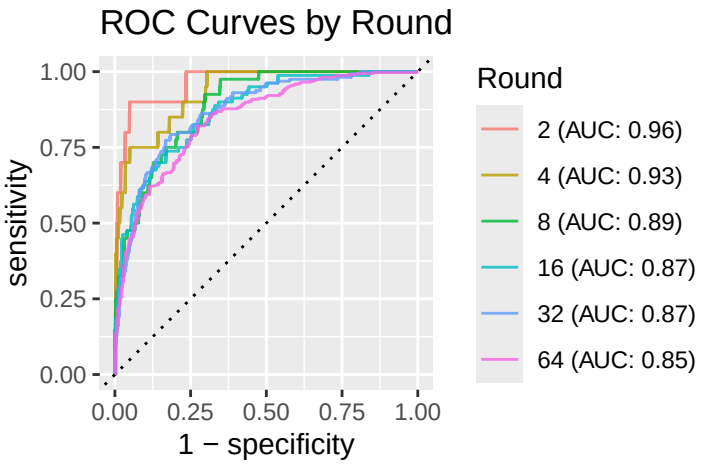
Side note on interpretation: our response is `ROUND`, a factor representing the round a team lost in, but a linear model complicates interpretation since it'll consider `ROUND` as numerical. Therefore, we'll consider a new output, the number of games that a team missed out on, which can be calculated as $\log_2$ `ROUND`. So, a team that lost in the Final Four would have an output of $\log_2 4 = 2$, where 2 represents the two games that they missed out on (the Final Four game and the championship). A team that wins it all would have an output of $\log_2 1 = 0$, as they didn't miss out on any games. This output can be thought of as a score that represents how well a team performed in the tournament: the closer you are to 0, the better the score.

Using this, our coefficients yield interesting interpretations. As `AdjOE` increases by 1 unit (as offense gets better), the model predicts `ROUND` to decrease by 16 holding other predictors constant, or for a team to advanced 2 games further. Equivalently, as `AdjDE` increases by 1 unit (as defense gets worse), the model also predicts `ROUND` to decrease, this time by 15 holding other predictors constant, or for a team to advance 3.9 games further. This seems strange at first, but must be interpreted alongside our interaction term. If we reframe the model so that improving defense corresponds to decreasing AdjDE (or increasing -AdjDE), we observe that the interaction between offense and defense is beneficial: as both improve together, we expect teams to advance further, with each unit improvement decreasing `ROUND` by 0.15 beyond what would be expected by adding their separate effects.

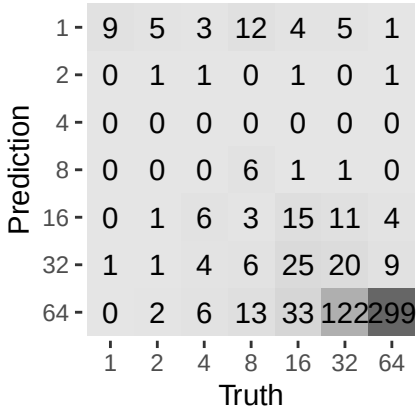
Overall, it seems that both `AdjOE` and `AdjDE` have significant effects on performance, but are trade-offs: it's difficult for teams to optimize defense and offense simultaneously. Notably, offense may be slightly more valuable, but as seen previously, this also largely depends on which round is being played in, as well as matchups.

Logistic

To adjust our thresholds and verify model performance, we used ROC curves for each of the 6 models. Our AUC is consistently solid, starting around 0.85 and increasing until a staggering 0.96, likely because there are fewer data points as we approach the championship game, but still indicating great performance.



Our goal is to maximize specificity to reduce false confidence: we don't want situations where we mistakenly predict a team will win but they actually lose. Therefore, we computed some new thresholds based on a minimum specificity of 95 %.



While an overall accuracy of 55.38% may not seem fantastic, we believe this is actually a decently high-performing model. Accuracy does not consider the many near-misses we had, where 122 32-rounders were predicted to finish in the round of 64 for instance.

Conclusion

We originally set to build models that could be better than random guessing or historical seed wins, and we believe that we've succeeded. Our models outperform random guessing significantly, (which require consecutive good guesses), and are likely similar to sophisticated seed methods. We also discovered multiple interesting trends in our analysis, notably that offense and defense seem to be trade-offs, and that offense matters more overall, with different effects depending on the round.

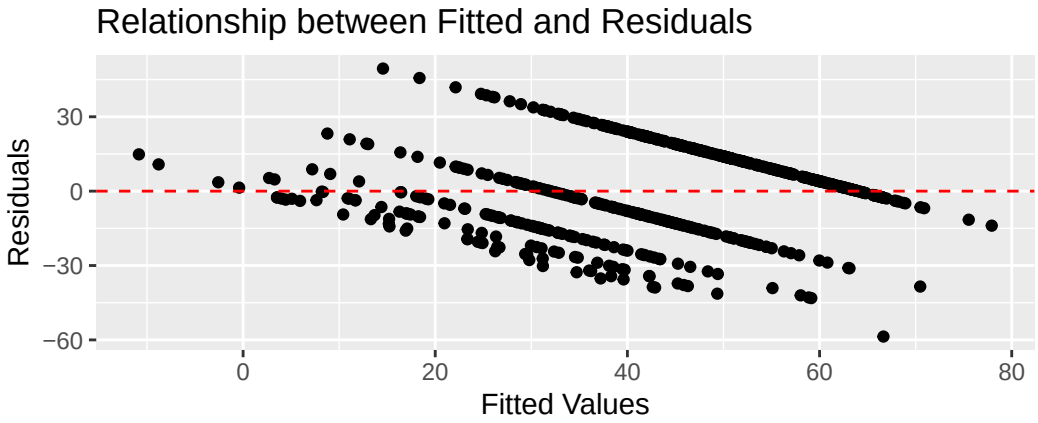
However, there were also a few limitations to our project. Our data timeframe is limited, meaning that we miss out on important seasons that could add more nuance to the model. This is especially important for our logistic system, where data is reduced immensely for our championship game, meaning that these models are likely overfit. There are also likely better ways to capture the effects of our other predictors without removing them through interactions that we weren't able to test. Furthermore, our linear assumptions were violated, likely due to our response `ROUND` being a factor, so more work could be done to deal with this kind of response variable.

Finally, there can definitely be further work done to improve our models. While we are able to predict in theory the round that teams would be able to reach, we could account for head-to-head battles where individual statistics are matched directly against each other instead of decoding the relationship of the statistics to the round reached. Additionally, more advanced statistical metrics may offer a new perspective on how to further provide insights, since the tournament is only played once annually and there is much volatility in team rosters as well as an evolution in playstyle.

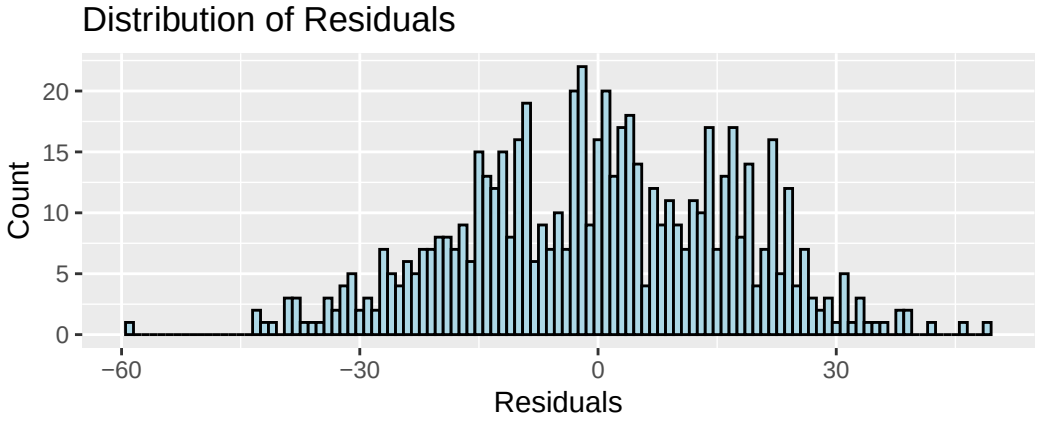
Taken as a whole, the beauty in the tournament resides in the fact that games are truly "madness", and as history suggests, we likely will never be able to fully predict a bracket with absolute confidence.

Appendix

Linear Assumptions

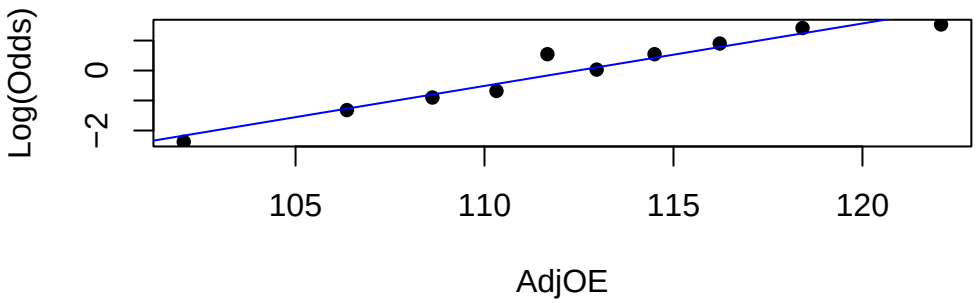


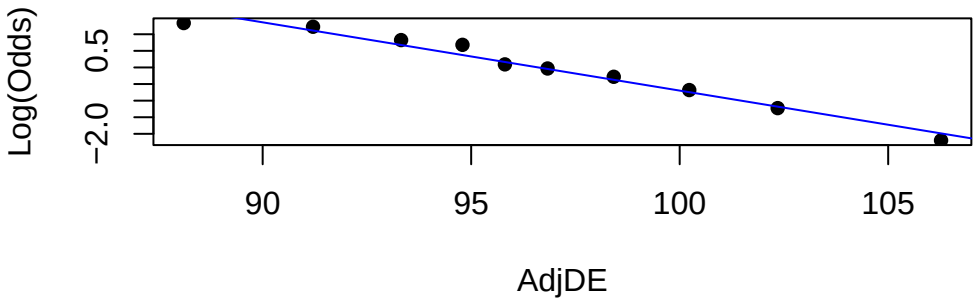
Linearity and constant variance seem to be violated in our residual plot, with diagonal lines forming across our residuals. This is likely because of the nature of our response as a factor (64, 32, 16, 8, 4, 2, 1). This violation persisted when we switched to log(ROUND), so we decided to keep it as such, still deriving value out of considering our response as numerical.



Our normality assumption is not terribly violated. There are regions that do not perfectly follow the curve, and some may argue that it is not uni-modal; regardless, our sample size is large so we assume the data to be robust to departures. Independence is a little tricky here, since teams are ranked based on performance against each other, but we considered our large sample size to be enough, with more data potentially being even stronger.

Logistic Assumptions





Our logistic model assumptions are a lot stronger. There is absolutely a linear relationship between the empirical logit and our predictors, meaning we might prefer logistic modeling to linear modeling in the context of our analysis. However, this assumption holds less as we get to later models with less data. Randomness and independence are a little tricky here, since teams are ranked compared to performance against each other. Regardless, we considered our large sample size across seasons to be enough, with even more data potentially being even stronger.