Regression Analysis of College Basketball Performance

Nick Reeder
Math 300 - Spring 2023

1. Introduction

The sports world is rapidly changing, innovations in statistics and sports medicine has pushed athletes to new levels and novel froms of social media have expanded the reach of competitions immensly. Because of this, a modern "money ball" ethic has developed requiring highly specified prediction and statistical analysis. College Basketball is a multi-billion dollar industry, with the governing body earning 1.15 billion dollars in 2021 alone [1]. This paper seeks to predict the success of a college basketball program using the Wins Above the Bubble (WAB), where the bubble refers to the wins required to earn a place at the NCAA Div 1 National Tournement (March Maddness). Two models were created. Model 1 was designed for coaches and institutions which can be used to increase performance and chances of success, and model two for sports bettors to predict with maximum accuracy and create the most efficient betting lines. This is an essential task as 15.5 billion dollars is expected to be wagered on this year's March Madness alone [2]. The focus on college basketball was chosen due to the size of the industry and the recent developments in the use of data science and statistics to improve performance [3].

2. Data and Model Creation

2.1 Data Description

The data was scraped from multiple sports statistics websites and coalated onto Kaggle by Andrew Sundberg. It came in a csv file 2455 observations tall by 22 variables wide, with each observation being the season performance of a team for one year. It is sampled from 8 years, 2013-2021, 355 teams, and 35 conferences. The predicted statistic, WAB, spans from -20 to 9. In order to maintain indpendence of observations and prevent psuedo-replication data for each team was averaged over the 8 years and year was removed. Wins and seed were removed aswell as they are direct indicators WAB and would not be available mid-season. Conference was also removed as the addition of 34 dummy variables would have rapidly violated observation:variable ratio

rule. A separate analysis of conference was later conducted.

2.2 Innital analysis

A preliminary analysis was conducted to ensure that there is a relationship between gathered statistics and WAB. Figure 1 shows adjusted offensive efficiency plotted against adjusted defensive effeiency with teams color coded based on whether or not they made the tournament more than they missed.

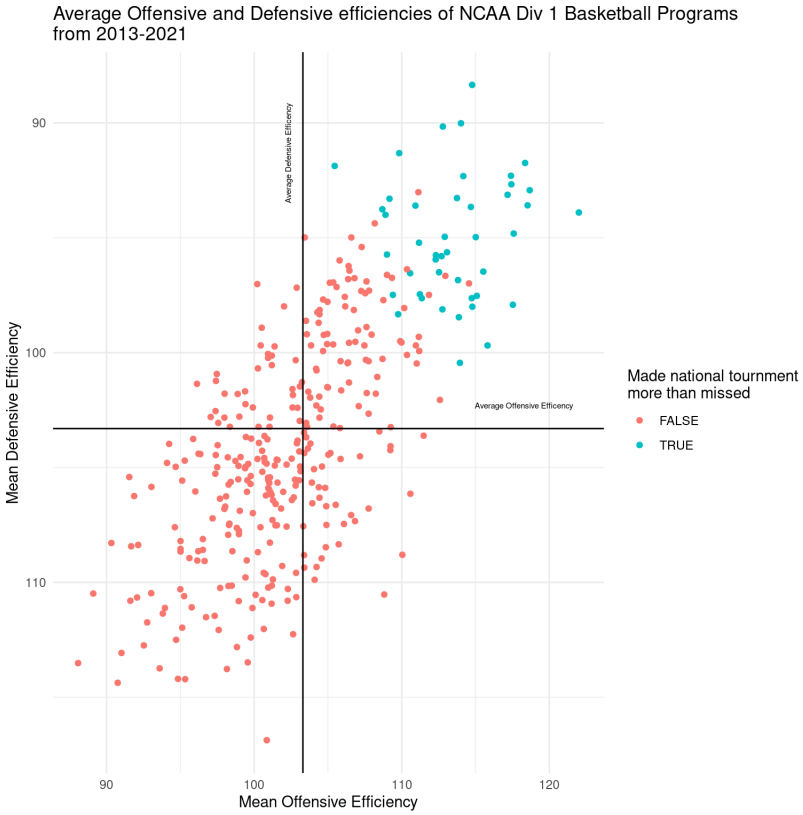

Figure 1: Average Offensive and Defensive Effecies plotted against eachother, axis lines are the overall averages

2.3 Variable Selection

After removing the 4 previous variables, the following 18 remained and are listed in Table 1.

In order to reduce the complexity of the model and improve parsimony, we performed a forwards and backwards stepwise regression. Both directions

Variable	Description
G	Games played that season
BARTHAG	Power Rating: Odds of beating an avg D1 team
ADJO	Adjusted Offensive Efficiency (Pts scored per 100 possessions)
ADJDE	Adjusted Defensive Efficiency (Pts allowed per 100 possessions)
EFG_O	Effective Field Goal Percentage made
EFG_D	Effective Field Goal Percentage allowed
TOR	Turnovers allows (rate)
TORD	Turnover Percentage forced (steal rate)
ORB	Offensive Rebounds
DRB	Offensive Rebounds allowed
FTR	Free Throw Attempt Rate
FTRD	Free Throw Attempt Rate Allowed
2p_O	2 point shooting percentage
2p_D	2 point shooting percentage allowed
3p_O	3 point shooting percentage
3p_D	3 point shooting percentage allowed
ADJ_T	Estimate of opposition's possessions per 40 mins

Table 1: Variable names and descriptions for all variables used in regression

agreed and removed the following terms: EFG-O, TOR-O, DRB, 2P-O. After looking at the summary output, FTR had a high p-value. Because of this we perfromed a nested f-test with and without FTR, and concluded that FTR was not statistically useful for prediction.

2.4 Expanding with higher order terms and model comparison

Once the base model was created, we performed a series of nested f-tests to add higher order and interaction terms. Interacting offensive and defensive effencies and offensive and defensive 3 point percentages yielded significant p-values, leading to their inclusion in the final model. No higher order terms yielded a significant p-value.

3. Model Analysis

3.1 Multicolinearity

At this point we decided to look for multicolinearity. This was done because

making analysis and inferences about individual variables would be very important were a coach to use the model to improve performances. To solve the multicolinearity problem another model was created where games, offensive efficency, defensive efficiency, field goal percentage, and defensive free throw rate were removed. This is reasonable as the offensive and defensive efficiencies and field goal percentages are summary statistics and field goal percentage describes a team's 3 point and 2 point success rate. However, were a betting service such as Bet MGM or a sports analysis service like ESPN or the Athletic to use this model, it would be much more important to be right than to remove multicolinearity. It was this divide in use that lead to the model being split, one that was slightly more accurate but with a multicolinearity problem, and one that was free of all multicolinearity but was slightly less accurate.

3.2 Residual Analysis and normality assumptions

Begining with the non multicolinear model, the model passed all tests of normallity and residual analysis. Figure 2 displays a residual and normal q-q plot for model one in which the residuals are randomly distributed above and below 0 and the normal q-q plot is linear. The figure demonstates that the normality assumptions hold.

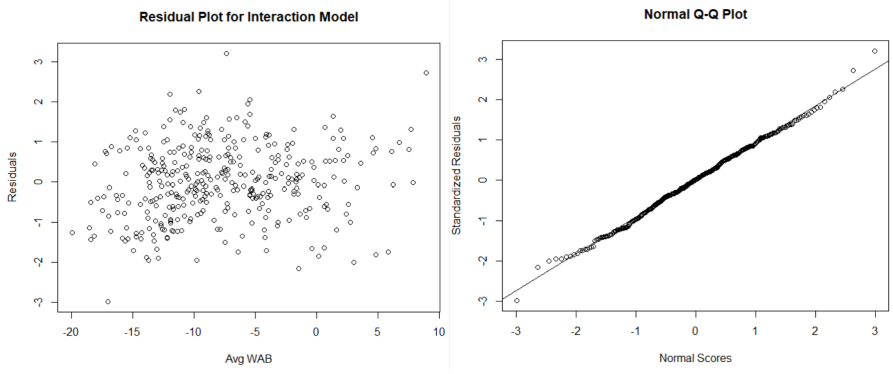

Figure 2: Residual Plot and Normal q-q for multicolinear model

This was also true of the multicolinear model, shown in Figure 3:

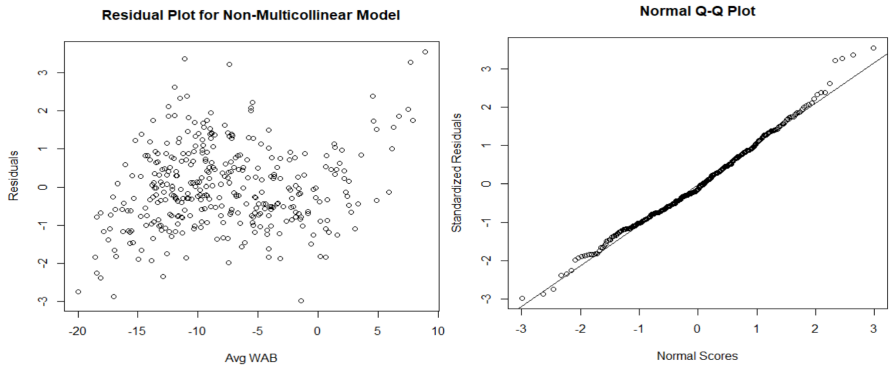

Figure 3: Residual Plot and Normal q-q for multicolinear model

3.3 Utility

Both models yielded very positve results for their utility statistics. Both model 1 and 2 demonstrated a significant global f-test with a p-value of 2.2^{-16}. Model 1 was able to account for 97.36% of error and model 2 was able to account for 96.51%, both very powerful adjusted r^2 values. Model 1 yielded an s and $2s$ of 0.9391 and 1.8782, respectively. This allows for high levels of confidence in the models accuracy. Model 2 yielded an s and $2s$ of 1.08 and 2.16, respectively. Despite being worse than model 1, this is still a very strong statistic, allowing for high confidence in the model's predictive power. Model 1 conclcuded with 12 variables and 14 beta parameters compared to model 2's 8 variables and 9 beta parameters. This is a reasonable number of estimaters given the volume of data that was fed into the model.

3.4 Train-Test Split

Given the high adjusted r^2 and s values, there were concerns of overfitting. To ensure that this was not the case a train-test split was performed on the mode using an 80-20 devide. The models performed very well, but not well enough to confirm overfitting. Model 1, the multicolinear model predicted WAB within 1 game at a 67.6% accuracy and within 2 games 97% accuracy. Model 2, the non-multicolinear model did slightly worse, prediting WAB within 1 game at 63.2% accuracy and within 2 games at 93% accuracy. This is in line with each model's s value.

4 Conference Analysis

Because conference was removed from the model, an analysis of conference was performed as it is reasonable to assume that some conferences would be better and would make it to the tournament more often. To do this, the distribution of National Championship Appearances by conference was compared to the uniform distrubtion via Chi-Square goodness of fit test. After computing the test, a p-value of 3.658^{-15} indicated that there was significant difference in the distributions, leading to the conclusion that there are some conferences, such as the ACC, that are far more likely to qualify for the National Chamoionship.

5 Concluson

In conclusion, it is clear that linear modeling techniques are an effective and powerful tool for the prediction of a teams ability to qualify for the tournament. This is important as statistic's value is becoming more and more valuable in sports. It is no longer relegated to the ESPN and CBS talking heads, but it actively being used in training and game day strategies. A more modern sports philosphy has evolved, where data and numbers drive decision making not belief or bias [3][4]. That said, sports statistics are not omnisciant and this is a good thing. 2023's truly mad March Madness has demonstrated that sports is often unpredictable. It is unlikely that any level of statistics would have been able to predicted the masterclass coaching of FDU's Tobin Anderson, or the massive underperformance from the 3 point line of the Alabama Chrimson Tide. As data grows in value in the sports world, so does its need for perfection.

References

[1]NCAA Div 1 Baskerball earnings

[2] American Gambling Ascociation 2023 March Madness report

[3] The use of data in UVA Basketball

[4] Moneyball: The Art of Winning an Unfair Game by Micheal Lewis