Jacob Thoma

# Modeling to Determine College Basketball Success: Which Stats Matter in College Basketball?

## ABSTRACT

Analytics are at the forefront of modern sports, responsible for generating plays, assisting in crucial decisions, and winning games. The relationship between wins, losses, and overall success of a college basketball program extends far beyond the simple statistics of a box score. This research aims to explain the role of statistics in regular season and tournament wins based on extensive Sports Reference and Kaggle data over the last five years of Division I College Basketball. Through methods of linear regression, logistic regression, stepwise selection, and data manipulation, this research paper details which advanced statistics have the greatest contribution to determining the regular season win-loss percentage and March Madness success. The findings of these methods suggest that stats related to generating possessions and capitalizing on created possessions are crucial to success in regular season and NCAA Tournament gameplay.

## INTRODUCTION

One does not need a statistics background, knowledge of basketball history, or even a computer to tell you that the most important stat in all of basketball is the one that wins games. When Caleb Love hit that pivotal three in the waning moments of the all-time classic UNC vs. Duke Final Four showdown last April, no one was worried about his tendencies for inefficient, off-balance jump shots or turnover bouts. Basketball, especially during March Madness, is a game of moments that is challenging to predict. However, analyzing a larger scale across modern college basketball history can provide insight into the statistics most responsible for creating pivotal tournament matchups, setting up historic shots, and crowning champions. With a combination of advanced statistics, quantitative analysis, and basketball knowledge, future college basketball success can be predicted through common stats found in Sports Reference (Sports Reference, 2023) and Andrew Sundberg's College Basketball Dataset, referenced here as the Kaggle Dataset (Sundberg A., 2021).

## MATERIALS AND METHODS

This study used data from the last five years of NCCA Division I matchups to identify variables that serve as potential predictors of team success in future seasons of similar playstyle.

The first and most important process to this analysis involved choosing which subsets of Sports Reference and Kaggle databases to consider. Only the 2017-2022 NCCA seasons were considered due to an assumed similarity in playstyle that would translate best for predicting future data. The reasons behind this are due to the overall evolution of game strategy during the 2010's with the increasing popularity of the three-point shot and the increasing three-point percentage across all five positions in the NBA that is assumed to translate over to NCCA Division I play. More information supporting this reasoning can be found here.

After the datasets were imported into R, initial issues with multicollinearity led to the decision of only using what Sports Reference considers "Advanced School Stats". Overall, the advanced statistics and respective opposing advanced statistics account for box score stats and extend upon them, leaving no reason to use the "Simple School Stats". In addition, to ease the overall process of comparing the data across variables that were measured in different units, a standardized data set was created with each data set.

Methods for analyzing the data included calculating correlations between variables and generating a summary of datasets to evaluate normality, pattern of residuals, and high leverage points. Stepwise selection, using RStudio's *leaps* package, was used to determine the best-fitting linear model that predicted regular season win percentage using the advanced statistics and the opponent's advanced statistics, creating the model denoted here as "bestlinmod". In addition, holdout and training datasets were used in a cross-validation process, with "bigtrainmod" denoting the extension of the training model to the entire dataset.

A new variable, called "tourney_team" was created as a binary indicator, where 1 indicates an NCAA tournament team and 0 indicates a non-tournament team. Logistic regression was run using the *bestglm* package. The *pscl* package was used to calculate model efficiency using McFadden's Pseudo R-squared (McFadden D., 1973). The "logmod" model is a logistic model that includes all advanced statistics and "logmod.3" is a logistic model that only includes Win-Loss percentage (WLper) and strength of schedule (SOS).

Lastly, the Kaggle dataset was used for its categorical variable “POSTSEASON” that indicated the furthest round an individual team went in the NCAA Tournament for a given year. POSTSEASON served as a non-binary categorical response variable that allowed for graphical comparison of each teams NCAA Tournament success using package *ggplot2*. Note that the only Kaggle data post-2016 with the POSTSEASON stat are the years 2017-2019.

## RESULTS

### Linear Regression Results

The training model from the cross-validation process recorded a shrinkage of -0.016, indicating that the training model predicted holdout data even better than the training data. “bigtrainmod” is an extension of the training model to the entire dataset (Figure 1). The stepwise selection model “bestlinmod” and the training model “bigtrainmod” had adjusted r-squared values of 0.8718 and 0.8716 respectively, indicating that each model accounts for an almost equal percentage of the variance in WLper. Both models have nine predictors and share seven in common, with marginal differences in the data set. The seven predictors shared between the models were all deemed statistically significant at the 0.1 significance level and are as follows: offensive rating (ORtg), opposing offensive rating (oORtg), free throw per field goal attempt (FTperFGA), opposing free throw per field goal attempt (oFTperFGA), opposing offensive rebound percentage (oORBper), opposing assist percentage (oASTper), and turnover percentage (TOVper).

The training model was selected to run analysis on the 2017-2022 Division I regular season data due to cross validation and its proximity in predictions to the best possible linear model. The residual diagnostics of the “bigtrainmod” are shown below in Figures 2, 3, and 4.

Figure 1. “bigtrainmod” summary

```
summary(bigtrainmod)
```

```
Call:
lm(formula = WLper ~ ORtg + oORtg + FTperFGA + oFTperFGA + oORBper +
    SOS + oASTper + TOVper + oBLKper, data = cbbadvS)

Residuals:
     Min       1Q   Median       3Q      Max
-1.46638 -0.24221  0.00754  0.23709  1.98873

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.716e-16  8.540e-03   0.000   1.0000
ORtg         5.830e-01  1.245e-02  46.834  < 2e-16 ***
oORtg       -5.623e-01  9.497e-03 -59.211  < 2e-16 ***
FTperFGA     7.473e-02  9.351e-03   7.991 2.41e-15 ***
oFTperFGA   -5.253e-02  9.225e-03  -5.694 1.45e-08 ***
oORBper      4.973e-02  1.006e-02   4.945 8.32e-07 ***
SOS          1.497e-02  9.226e-03   1.623   0.1048
oASTper     -2.409e-02  9.613e-03  -2.506   0.0123 *
TOVper      -2.189e-02  1.153e-02  -1.899   0.0577 .
oBLKper     -6.607e-03  8.920e-03  -0.741   0.4590
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3584 on 1751 degrees of freedom
Multiple R-squared:  0.8722,    Adjusted R-squared:  0.8716
F-statistic:  1328 on 9 and 1751 DF,  p-value: < 2.2e-16
```

Figure2. “bigtrainmod” residuals vs. fitted

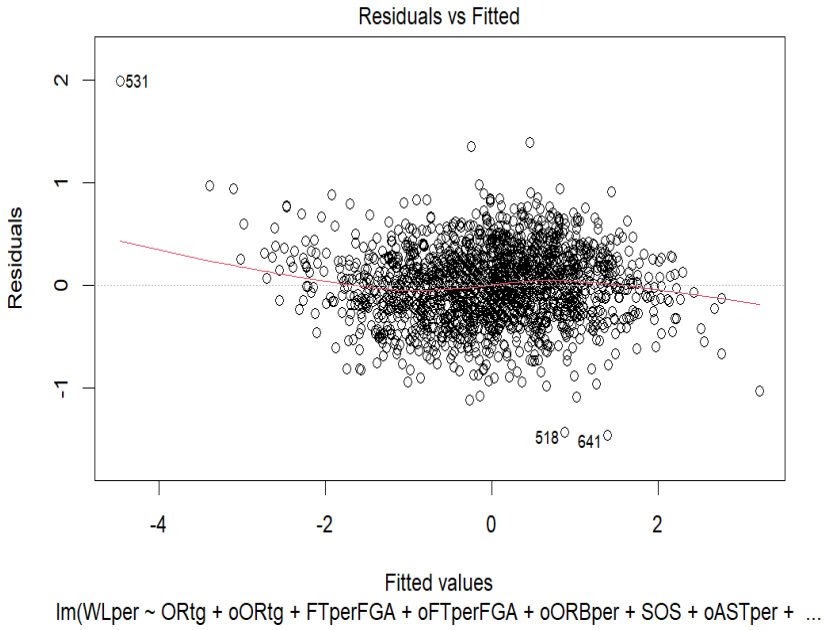



Figure 3. “bigtrainmod” normality

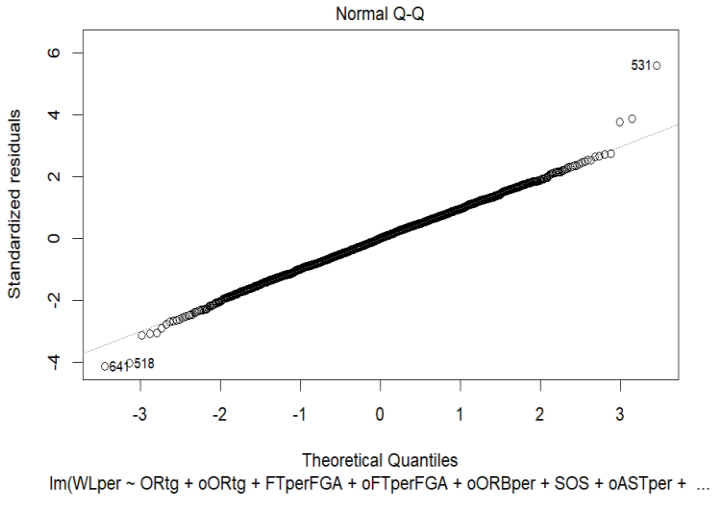
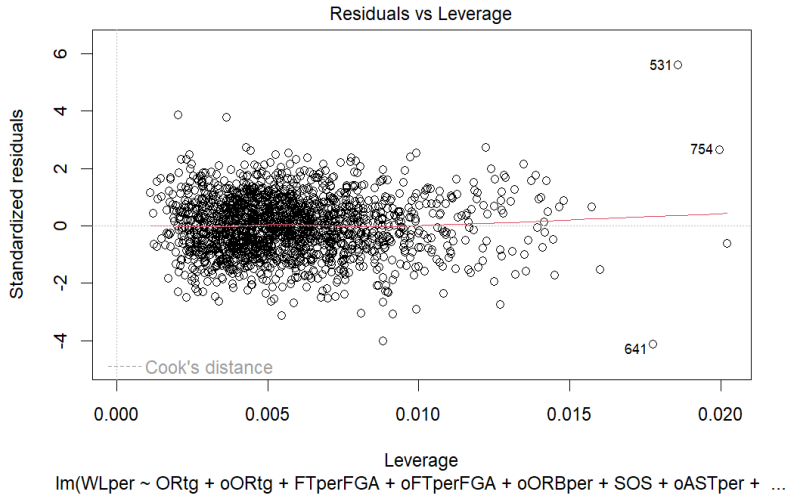



Figure 4. Residuals vs. Leverage



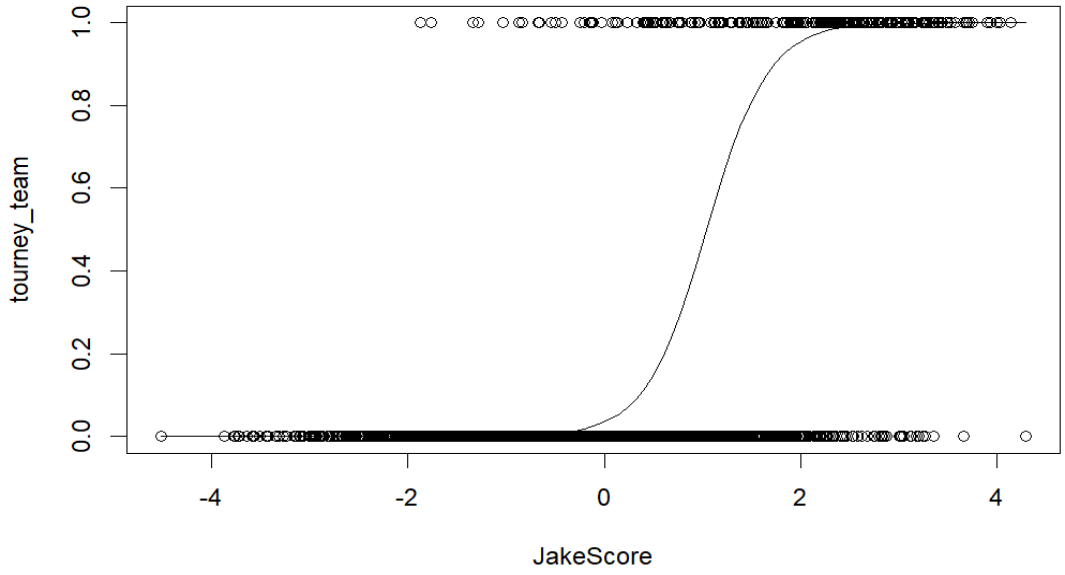

The diagnostic plots of “bigtrainmod” yielded interesting results (Figures 2, 3 and 4). Besides several influential points, the diagnostic plots above indicate that the linear regression model meets all the assumptions of linear regression. The high leverage points 531 and 754 are cases where the model overpredicted team success due to these teams’ historical ineptitude in categories not included in the model, such as true rebounding percentage and turnover percentage that drastically change a team’s number of possessions whereas influential point 641

was underpredicted due to the drastic success in categories that generated possessions such as steal percentage that were not considered in “bestlinmod”.

Logistic Regression

The logmod and logmod.3 logistic regression models were compared, and their McFadden Pseudo R-squared values were 0.45 and 0.43 respectively. A Drop in Deviance test was performed between logmod and logmod.3 that did not find statistically significant differences between the logistic models. A linear combination of WLper and SOS, denoted here as “JakeScore”, was used for a rough visualization of the logmod.3 logistic regression model, as shown in Figure 5. Note that “Jake Score” has a similar logistic fit but should not be confused with logmod.3

Figure 5. “Tourney_team vs. JakeScore”




Kaggle Graphical Analysis

The POSTSEASON categorical response variable was analyzed using graphical analysis of the Kaggle dataset. The relationship between adjusted offensive efficiency (ADJOE) and allowed adjusted offensive efficiency (ADJDE) is measured in Figures 6 and 7. Figure 6 is a scatterplot comparison where data is separated categorically by furthest round in the NCAA Tournament. Figure 7 provides a visualization of the differences between ADJOE (yellow) and ADJDE (purple) teams that made it to each stage of the NCAA Tournament.

Figure 6. “ADJOE vs. ADJDE”

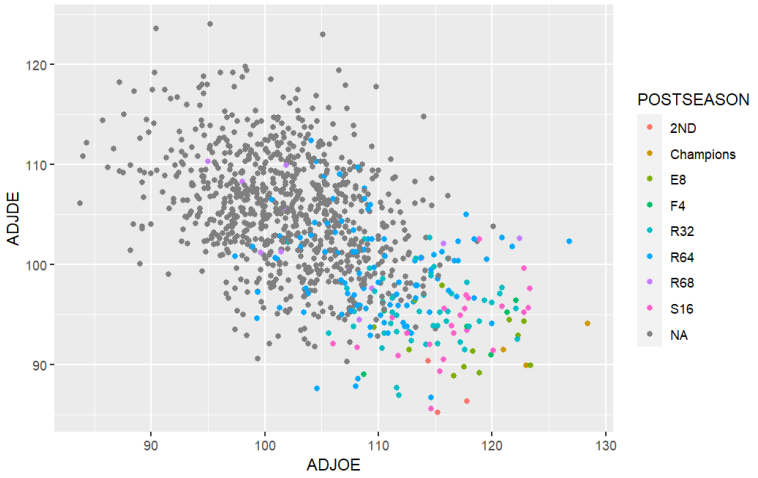
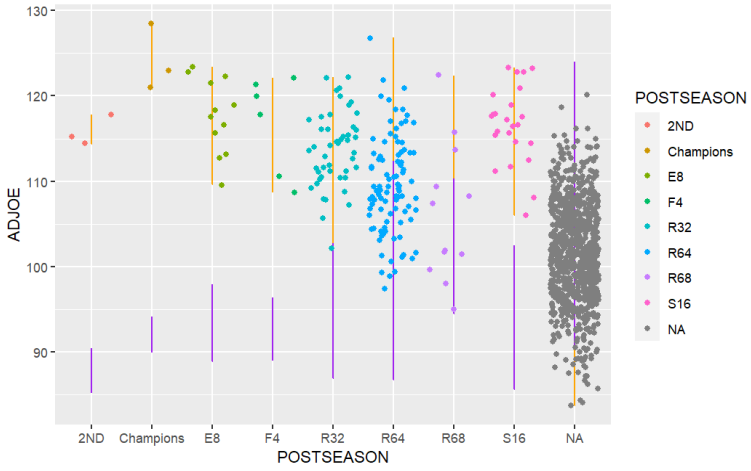



Figure 7. “ADJOE vs ADJDE by POSTEASON”




## DISCUSSION

### Linear Regression

The linear training model used for cross validation not only modeled holdout data accurately, but also included seven of the nine predictors found using a stepwise selection process across the entire model. The first two predictors included in the linear model are a team’s offensive rating (ORtg) and opponent offensive (oORtg). These two stats measure the number of points scored or scored against per 100 possessions, respectively. Intuitively, being able to outscore an opponent per 100 possessions on average will separate successful NCAA regular season teams from losing programs and thus dictate the importance of ORtg and oORtg in the linear model predicting WLper. However, the number of possessions per team each game is variable. As such, it makes sense that stats such as opposing offensive rebound percentage (oORBper) and turnover percentage (TOVper) are included in both linear models to adjust for the variability in the number of possessions per team per game.

The four stats above account for variability in pure offensive production and offensive production allowed per 100 possessions, but there are other important predictors at play. oFTperFGA and FTperFGA were also considered statistically significant predictors across both linear models as they account for points generated from shooting fouls. Shooting fouls are plentiful in basketball, and the ability for a team to get to the free throw and convert free throw attempts is critical to offensive success and therefore critical to determining win loss percentage.

The same reasoning applies to oFTperFGA, as committing a consistently high number of shooting fouls per game is a recipe for a very efficient opposing offense on a per game basis.

The last statistically significant stat found in both linear regression models is the opposing assist percentage against a team (oASTper). This stat was likely included in the linear regression model to penalize WLper predictions for teams with bad overall team defense and transition play, which are crucial to the outcomes of basketball games but not necessarily quantified by the other stats included in the linear regression model.

Logistic Regression

The logistic regression model using all variables and the model using only WLper and SOS were found to have no statistically significant difference in predicting admission to the NCAA tournament. The WLper and SOS logistic regression model was thus chosen for simplicity and evaluated at a McFadden Pseudo R-squared score of 0.43, indicating an excellent fit of the data to a logistic regression model (Hemmert 2016). Due to the nature of NCAA Tournament admission procedures, is also intuitive that WLper and SOS are the main predictors of NCAA Tournament status. Each of the 32 NCAA Division I Conference Champions receives an automatic bid to the NCAA Tournament, while the other 36 receive an at large bid from the selection committee. While there is some considerable variability in the Conference Champions per conference based on these conference tournaments, the selection committee's ability to select the last 36 teams in the tournament is primarily with the consideration of a team's WLper and SOS and explains why these specific stats are essential in predicting bids to the NCAA tournament.

Kaggle Graphical Analysis

Figures 6 and 7 show two representations of the relationship between adjusted offensive efficiency (ADJOE), adjusted offensive efficiency allowed (ADJDE), and the furthest round a team went in the NCAA tournament each year between 2017 and 2019. ADJOE is defined as the average number of points a team would score per 100 possessions on the average D-1 defense and ADJDE is the average points a team would allow to the average D-1 offense (Sundberg 2021). The separation between categories of teams with different NCAA tournament results (each represented by different colors) is evident. On average, teams with the better combination

of high ADJOE and low ADJDE tend to make it further in the March Madness brackets. This graphical analysis contributes to the importance of completeness from an individual team perspective.

Even teams with the most exciting, prolific, and efficient offenses cannot go far in the tournament without at least mediocre ADJDE. A vast majority of ‘Sweet Sixteen’ or better teams had an ADJDE of less than 97, and even teams with a very high ADJOE tended to fall short in the later tournament stages if their ADJDE was below this threshold.

However, teams with a stifling defense that did not have the offensive efficiency to capitalize on forced turnovers and possession disparities did not achieve ultimate success in the NCAA Championship. The second-place teams within this graph evidence this point well, as all three of which had an impressive ADJDE but low ADJOE values relative to most Sweet Sixteen or better teams.

From all angles of “college basketball success”, it is evident that possession-based stats and efficiency-based stats are the driving factors for regular season success, tournament bids, and March Madness success. In a simple sense, the answer to “Which stats matter in college basketball?” is the question “Which teams can get the ball the most, and do the most with it?”. Now we have a question that the data can answer!

# References

Hemmert, G. A., Schons, L. M., Wieseke, J., & Schimmelpfennig, H. (2016). Log-likelihood-based pseudo-*R* squared in logistic regression. *Sociological Methods & Research, 47*(3), 507–531. https://doi.org/10.1177/0049124116638107

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. *University of California Berkeley*.

Sports Reference. (2023). NCAA Seasons Index [Database]. Retrieved 2022, December 22, from SRCBB.

Sundberg, A. (2021). College Basketball Dataset [Database]. Retrieved 2022, December 22 from https://www.kaggle.com/datasets/andrewsundberg/college-basketball-dataset