Relationship Between NCAA Men's Basketball Non-Game Statistics and Team Success

Alexander Albrecht and Amélie Devine

STAT 4315: Applied Statistical Models

Villanova University

Professor Alcantara

December 11, 2023

A. Introduction

With the changing landscape of NIL in the college basketball arena, success from college basketball programs has continued to vary substantially from year to year. Even so, there is a general trend of continual success from storied programs. However, there is a known disparity in the overall wealth of schools and the amount of money that they spend on their programs, a distinction made even more significant considering NIL. Additionally, with the recent departures of many headlining programs from longstanding conferences, understanding the relation between a school's revenues (much of which is driven by conference-wide media deals) many other off-the-court factors, and a school's success in basketball is valuable information to understand the legitimacy of these conference changes. With this project, we hope to get a better understanding of off-the-court factors that may impact a men's college basketball team's success. We are particularly interested in seeing which variables can be considered significant predictors and if any of them are strongly correlated with a team's success. So, our research questions for this project were as follows: Do off-the-court factors affect a school's basketball results for an individual season? If so, which factors are most important in predicting the results of a season? We hope to find a model that accurately shows this relationship, as well as which predictors are significant and/or correlated. There is not much existing research on how the revenues, expenses, region, enrollment, or arena capacity affect winning. Therefore, we had little idea of what to expect.

Our data encompasses 96 Division I colleges from the so-called "Power" conferences that tend to dominate men's college basketball. These conferences are the ACC, Big 12, Big East, BIG-10, Mountain West, PAC-12, SEC, and WCC (listed 1-8 in Conf_Cat). The explanatory variables are conference, years in conference, enrollment, public/private, football revenue (in

millions of US dollars), men's basketball revenue (in millions of US dollars), men's basketball expenses (in millions of US dollars), grand total revenues (in millions of US dollars), and men's basketball arena capacity. Our response variable is a school's winning percentage in the 2021-2022 season.

B. Exploratory Data Analysis

Our original data set included 53 different columns of information on each school, including information about athletic directors, media rights, football, basketball, and non-money-making sports' monetary information. Initially, we decided to look at how 11 variables affect winning percentage. These variables, which were also discussed above, were Conference (Categorical), School State (Categorical), Years in Conference, Enrollment, public/private (categorical), football revenue, men's basketball revenue, men's basketball expenses, men's basketball profit, grand total revenues, and men's basketball arena capacity. But, after considering the variables, we knew we needed to make a couple of changes. First, we needed to assign a category number to each conference in order to make it a valid categorical data set. This is demonstrated below.

Conferences & Categories	
ACC	1
Big 12	2
Big East	3
BIG-10	4
Mountain West	5
PAC-12	6
SEC	7
WCC	8

Then, we realized that school state may not actually have a big impact on our analysis and instead decided to convert states into regions (as determined by the US government). We assigned each region a numerical value as well, as seen below.

Region	
West	1
South	2
Midwest	3
Northeast	4

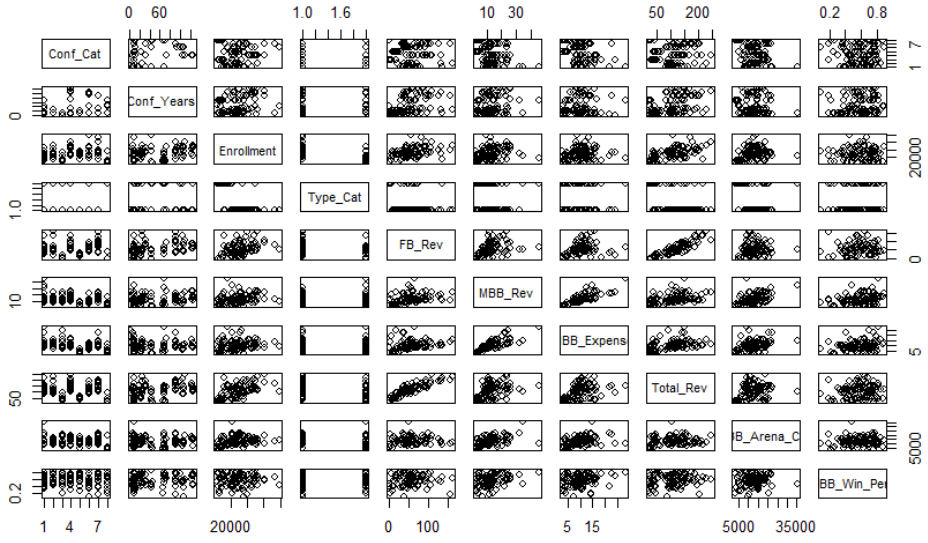
We ultimately decided to remove Region from our analysis after later consideration. Then, we decided to convert Public and Private's categorical data to two different numerical values, as seen below.

University Type	
Public	1
Private	2

Once we obtained a cleaner dataset, we had to consider which categories would be helpful to analyze from here. We decided to remove MBB Revenue-Expenses due to its plausible multicollinearity with MBB Revenue and MBB Expenses. We also converted all monetary values to amounts in millions of dollars to make model building easier to interpret. With all of this in mind, we began working on our model building, including our initial perception that Football Revenue and MBB Revenue would be the two most important variables in MBB success.

```
   Conf_Cat       Conf_Years       Enrollment       Type_Cat         FB_Rev          MBB_Rev
Min.   :1.000   Min.   :  2.00   Min.   : 3200   Min.   :1.000   Min.   :  1.235   Min.   : 2.690
1st Qu.:2.000   1st Qu.: 10.00   1st Qu.:17593   1st Qu.:1.000   1st Qu.: 35.850   1st Qu.: 8.458
Median :4.000   Median : 31.00   Median :27950   Median :1.000   Median : 52.242   Median :12.227
Mean   :4.417   Mean   : 48.98   Mean   :28946   Mean   :1.323   Mean   : 58.358   Mean   :13.247
3rd Qu.:6.250   3rd Qu.: 90.00   3rd Qu.:36947   3rd Qu.:2.000   3rd Qu.: 78.697   3rd Qu.:16.262
Max.   :8.000   Max.   :126.00   Max.   :79232   Max.   :2.000   Max.   :161.533   Max.   :45.109
                                                                 NA's   :15
 MBB_Expense       Total_Rev      MBB_Arena_Cap    MBB_Win_Perc
Min.   : 2.690   Min.   : 20.78   Min.   : 3104   Min.   :0.0968
1st Qu.: 7.594   1st Qu.: 48.84   1st Qu.: 9454   1st Qu.:0.4801
Median : 9.992   Median :108.46   Median :12494   Median :0.5942
Mean   :10.343   Mean   :106.63   Mean   :12817   Mean   :0.5802
3rd Qu.:12.416   3rd Qu.:144.20   3rd Qu.:15390   3rd Qu.:0.7154
Max.   :28.020   Max.   :246.61   Max.   :35446   Max.   :0.8919
```

In the above figure, we see a summary of all of the different explanatory variables. There was a high amount of variability in football revenue, ranging from 1.2 million to 160 million, along with 15 schools not having a football program and thus no revenue.



The above graph shows a slight upward trend, so a slight positive correlation between the variables of men's basketball revenue and men's basketball win percentage and football revenue and men's basketball win percentage. After initial analysis, we decided to convert all monetary columns' values into millions of dollars since the amounts were so minuscule when they were in their original amounts.

C. Regression Analysis

We began by running a linear regression on all of the variables and checking the summary statistics of the slopes. We noticed that only the intercept and only one of our variables - Men's Basketball Revenue were significant at the 0.05 level. So, we decided to run a few different model selection techniques to pick a better, reduced model. We started with Best Subset Regression, which selected Model 1 (with Men's Basketball Revenue as the sole predictor) as the best model, since it fits 4/6 optimality criterion, with the highest R^2_{adj} and R^2_{press} and lowest AIC and BIC. Then, we ran both Stepwise Regression and Backward

x	r2	adjr2	AIC	BIC	r2press	Cp
MBB_Rev	0.1029232	0.093379792	-73.39222	-65.69918	0.06739601	-4.2247161
Type_Cat MBB_Rev	0.1091770	0.090019531	-72.06382	-61.80643	0.05015541	-2.8366298
Conf_Cat Type_Cat MBB_Rev	0.1122669	0.083319121	-70.39739	-57.57564	0.03465980	-1.1389661
Conf_Cat Conf_Years Type_Cat MBB_Rev	0.1130322	0.074044600	-68.48018	-53.09409	0.01385079	0.7861566
Conf_Cat Conf_Years Type_Cat MBB_Rev MBB_Expense	0.1133328	0.064073514	-66.51272	-48.56228	-0.01340984	2.7567434
Conf_Cat FB_Rev MBB_Rev MBB_Expense Total_Rev MBB_Arena_Cap	0.1137947	0.041940162	-54.29030	-35.13471	-0.07828786	4.1010231
Conf_Cat Type_Cat FB_Rev MBB_Rev MBB_Expense Total_Rev MBB_Arena_Cap	0.1150331	0.030173247	-52.40358	-30.85354	-0.12876150	6.0016622
Conf_Cat Conf_Years Type_Cat FB_Rev MBB_Rev MBB_Expense Total_Rev MBB_Arena_Cap	0.1150538	0.016726421	-50.40547	-26.46098	-0.16017053	8.0000022
Conf_Cat Conf_Years Enrollment Type_Cat FB_Rev MBB_Rev MBB_Expense Total_Rev MBB_Arena_Cap	0.1150538	0.002877527	-48.40547	-22.06653	-0.18417435	10.0000000

9 rows | 1-1 of 7 columns

Elimination, which both selected

Model 1 as the best model. We also tested our full model versus the reduced model, and it produced an F-statistic of 1.394, which is less than the F critical value of 2.048, so we failed to reject the null hypothesis and do not have evidence to show the full model is adequate. Through all of these tests and methods, we select a new linear fit model in R using Simple Linear Regression. This new, reduced model (in the image below) is as follows: Win Percentage = 0.48249 - 0.007378(MBB Revenue).

```
Call:
lm(formula = MBB_Win_Perc ~ MBB_Rev, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.44098 -0.10004  0.00514  0.13058  0.28528

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.482490   0.034033  14.177  < 2e-16 ***
MBB_Rev     0.007378   0.002247   3.284  0.00144 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1617 on 94 degrees of freedom
Multiple R-squared:  0.1029,    Adjusted R-squared:  0.09338
F-statistic: 10.78 on 1 and 94 DF,  p-value: 0.001438

Analysis of Variance Table

Response: MBB_Win_Perc
           Df  Sum Sq  Mean Sq F value   Pr(>F)
MBB_Rev     1 0.28204 0.282045  10.785 0.001438 **
Residuals 94 2.45830 0.026152
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

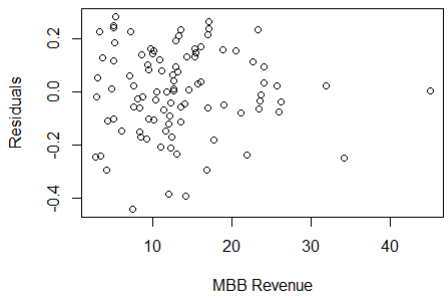
This tells us that if a school had $0 revenue from men’s basketball, they would be expected to win 48.25% of their games. Additionally, it says that as men’s basketball

revenue increases by $1 million, the winning percentage is expected to increase by as much as 0.7%. Our hypothesis test for the significance of the estimates gives a p-value of 0.002, which indicates that our overall model is significant. Our adjusted R^2 value for our model is .09338, indicating that only about 9.3% of a team's winning percentage can be explained by their MBB revenue.

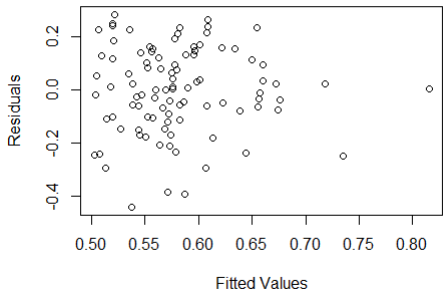
We used Villanova as a sample prediction. Using our model, Villanova was predicted to have a 0.6338 win percentage, but in 2021-2022, Villanova had a 0.7895 win percentage, which gives us a 19.7% error. However, that actual value does fall right within the prediction interval and not too far outside the confidence interval. The confidence interval allows us to conclude that we are 95% confident that the true increase in winning percentage per $1 million increase in men's basketball revenue is between 58.78% and 67.99%. Additionally, the prediction interval states that we are 95% confident that the true winning percentage of an NCAA men's basketball team with $21,513,366 revenue due to men's basketball is between 30.95% and 95.82%. Since we were predicting the value of a school, we looked to the prediction interval for the accuracy of our prediction, and it was entirely within expected.

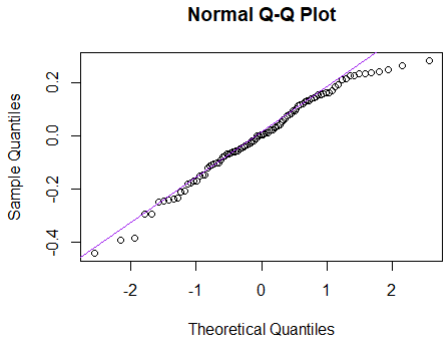
D. Model Diagnostics

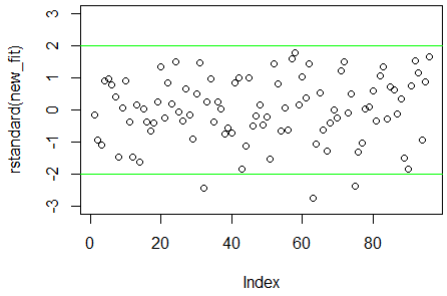
Next, we found it critical to check the assumptions underlying our regression model. To check for linearity, we plotted the residuals versus actual data points, which only required one graph since our model ended up being a simple linear regression. The plot showed the points to be randomly scattered around 0 with no obvious,

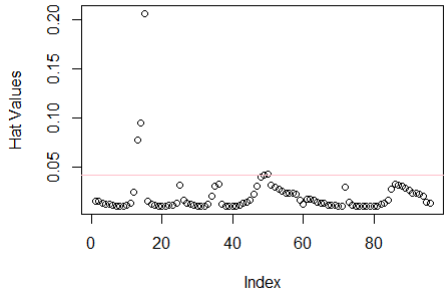


systematic pattern, so the linearity assumption is satisfied. To check for equal variances, we plotted the fitted values against the residuals, which also showed no systematic pattern, so constant variance is satisfied. We then created the QQ-Norm plot to check for normality, and the residuals seemed to fit the QQ line very well around the center, but varied a bit at either end, so we decided to run the Wilk-Shapiro Test to double-check normality. The Wilk-Shapiro Test had a p-value of 0.099, which means we fail to reject the null hypothesis that the data is normally distributed. This allows us to conclude that the normality assumption is satisfied. We then checked for outliers by plotting the standard deviations and only found 3 that fell outside the 2 standard deviations from the mean – which we found to be points 32, 63, and 75 by checking the standard deviations of all the data points using *rstandard* in R. Then, we wanted to check for influential points, so we plotted the hat values and compared it to the benchmark value of 0.0417 (based on predictors and number of data points). We found about 5 high influence points, so we decided to check the specific hat values and remove the highest influence point (15) to see what would happen.









When we replotted the hatvalues of the data without point 15, there were still 5 high influence points, so we decided to remove all 5 (13, 14, 47, 48, 49) and recheck the influence once more. We then had a lot more influential points, so we went back to the model and compare the summary statistics (slopes and significance) of our model with and without the influential points. The slopes were similar enough that it was clear the influential points did not have that much of an impact on the data, so we decided to keep those points in. Also, there were no points that were both outliers and high influence, so we did not need to remove any of them.

E. Summary and Conclusion

Let's reconsider our initial research questions of: "Do off-the-court factors affect a school's basketball results for an individual season? If so, which factors are most important in predicting the results of a season?" Our linear regression indicates that it is possible to predict the results of a season through off-court values. Our model is Win Percentage = 0.48249 - 0.007378(MBB Revenue). However, we cannot predict a team's winning percentage with much precision. Our adjusted R^2 value for the linear regression model is only about 9.3%, which means we cannot make too many significant claims about the relation between MBB revenue and MBB winning percentage. We can say that a team's men's basketball revenue and their winning percentage are lightly linearly correlated. Our tests indicated that no other variables were able to predict men's basketball winning percentage with any significance.

To improve our study, we may want to look at time-series data to see how revenues over time are correlated to a team's success over the years. We may also want to incorporate more

variables and smaller schools (schools that are in smaller conferences with less money) in order to help those schools with tighter budgets. We hope to continue to investigate this relationship with additional skills and model capabilities in the future.

Works Cited

"2021-2022 Division 1 Athletics Information." *Equity in Athletics*, ope.ed.gov/athletics/#/. Accessed 11 Dec. 2023.