What makes a good basketball team according to statistical data?

Chris Zhu & Ziyu Zhao
Denison University, Professor Zhe Wang

Objectives

we will examine the methodology of the ranking and have a more comprehensive viewpoint of how statistical variables affect college basketball teams in the percentage of winning. Also to discover the relationships among variables.

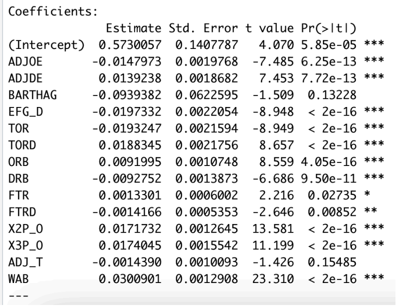
Introduction

The data is collected from division 1 college basketball teams in the US and has a collection of different statistical values that are collected during games. The dataset covers 353 basketball teams in 27 different conferences and 22 variables.We chose this topic because many people follow college basketball, and basketball teams want to be the best. So through our project we will find what a team needs statistically to be the best division 1 college basketball teams, and to find out how numeric and categorical data affect basketball teams.

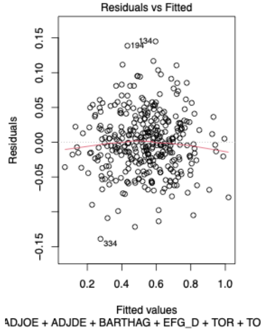
Methodology

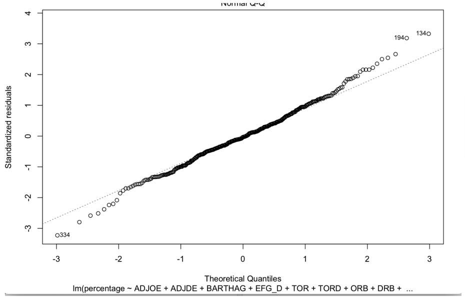
First, we conduct Step Akaike Information Criteria for all variables to see how each of the variable affects the percentage of winning games. We can see the remaining variables, which are more significant to affect the percentage of winning games. Then we check the p-value of each remaining variable. We conduct simple linear regression models for variables with p-value less than 0.05. Also run correlation tests and t-test for the variables.

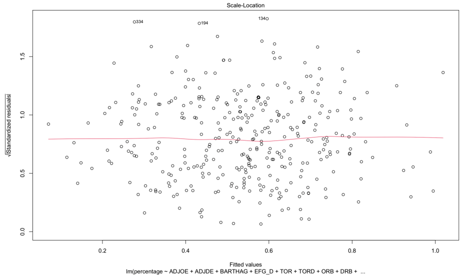
Then we use the ANOVA test to compare the group mean among remaining variables.



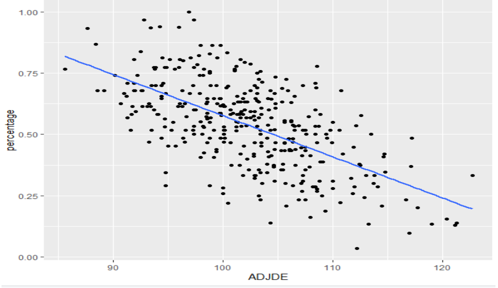
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.5730057  0.1407787   4.070 5.85e-05 ***
ADJOE       -0.0147973  0.0019768  -7.485 6.25e-13 ***
ADJDE        0.0139238  0.0018682   7.453 7.72e-13 ***
BARTHAG     -0.0939382  0.0622595  -1.509  0.13228    
EFG_D       -0.0197332  0.0022054  -8.948  < 2e-16 ***
TOR         -0.0193247  0.0021594  -8.949  < 2e-16 ***
TORD         0.0188345  0.0021756   8.657  < 2e-16 ***
ORB          0.0091995  0.0010748   8.559 4.05e-16 ***
DRB         -0.0092752  0.0013873  -6.686 9.50e-11 ***
FTR          0.0013301  0.0006002   2.216  0.02735 *  
FTRD        -0.0014166  0.0005353  -2.646  0.00852 ** 
X2P_O        0.0171732  0.0012645  13.581  < 2e-16 ***
X3P_O        0.0174045  0.0015542  11.199  < 2e-16 ***
ADJ_T       -0.0014390  0.0010093  -1.426  0.15485    
WAB          0.0300901  0.0012908  23.310  < 2e-16 ***
---
```







We also use the ggplot to show the relationship between ADJDE(Adjusted defensive efficiency)and winning percentage.



Results

Through the linear regression models and t-test results, we find the 14 out of 22 variables are significant for a team to win games and 12 of them with p-value less than 0.05. Linear regression models indicate how much remaining variables affect the percentage of winning games. Adjusted defensive efficiency and percentage of winning games are highly correlated and have a normal distribution. Residual vs Fitted plot indicates that the residuals and the fitted values are uncorrelated, as they should be in a homoscedastic linear model with normally distributed errors. Fitted values vs standardized residuals plot also shows there is no correlation between variables QQ plot suggests this is a near normal distribution.

Conclusion

We can include that 12 out of 22 variables are statically significant and have significant impact on the percentage of winning. And the 12 variables are the most important to form a good basketball team. The p-value and standard error are fairly small, so we can trust the results. Residual vs Fitted and fitted values vs standardized residuals plot makes sure variables do not affect each other to ensure the accuracy of the model.

References

Big 10 for best conference and ACC
https://www.ncaa.com/news/basketball-men/article/2020-11-09/college-basketball-rankings-complete-ap-top-25-instant-reaction-analysis