BIOS 611 Project 1

Jane She

e-mail: jane.she@unc.edu

19 October 2021

1 Introduction

Men's college basketball is a widely popular sport, especially at the Division 1 level and is broadcasted nationally for fans to watch. According to Cav's Corner, the UVA men's basketball team is one of the top revenue generators for the school.

As a recent graduate of the University of Virginia, I was interested in looking at various game statistics and their infuences on the number of games won by a team. I specificially looked at the 2019 NCAA D1 basketball season since that is both the year that UVA won the national championships and the last pre-COVID season.

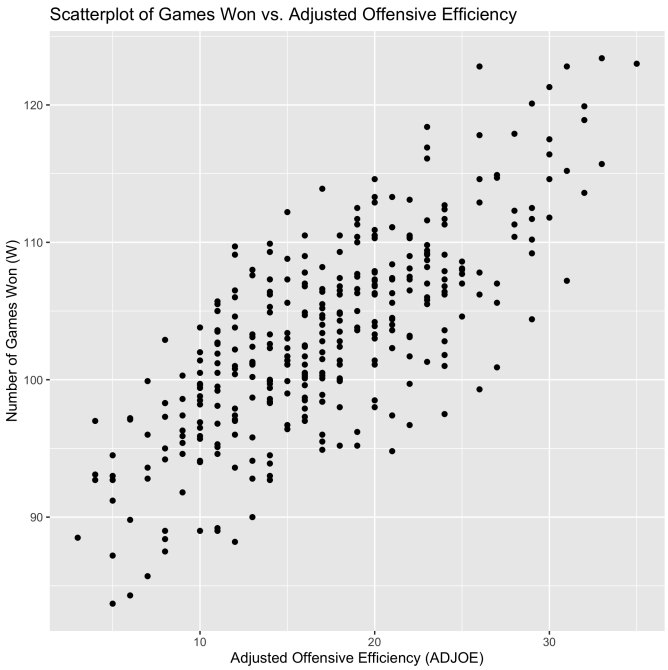
In this report, I examined the relationship between the number of games won and statistics including adjusted offensive efficiency, adjusted defensive efficiency, effective field goal percentage, turnover rate, and steal rate.

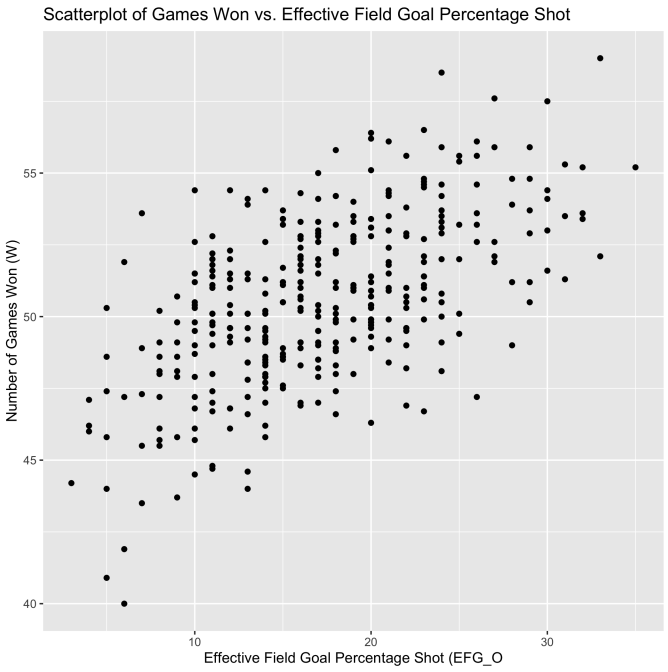
2 Data

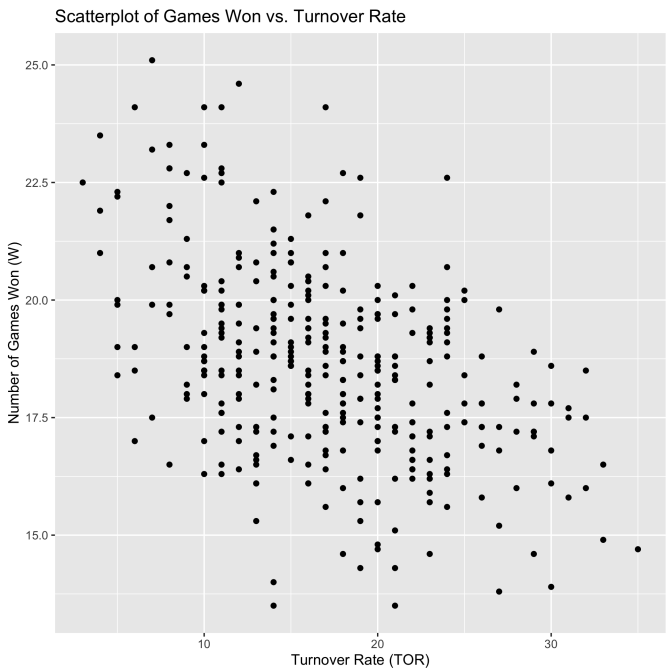
The dataset with variable descriptions and more context can be found here and was found on kaggle. Originally, the data was scraped from barttorvik.com, a famous college basketball website.

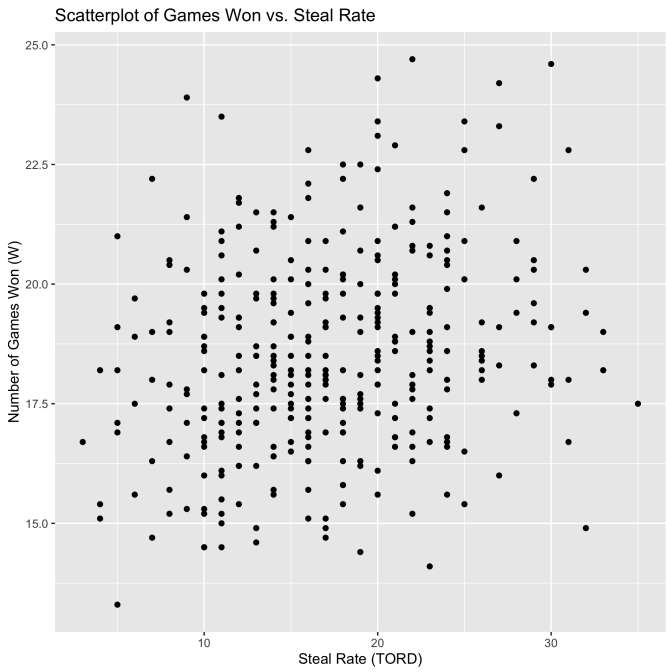
3 Exploratory Data Analysis

In our exploratory data analysis, I decided to choose a couple variables I was interested in and plot them against the number of wins. The variables I included are adjusted offensive efficiency, adjusted defensive efficiency, effective field goal percentage shot, turnover rate, and steal rate.



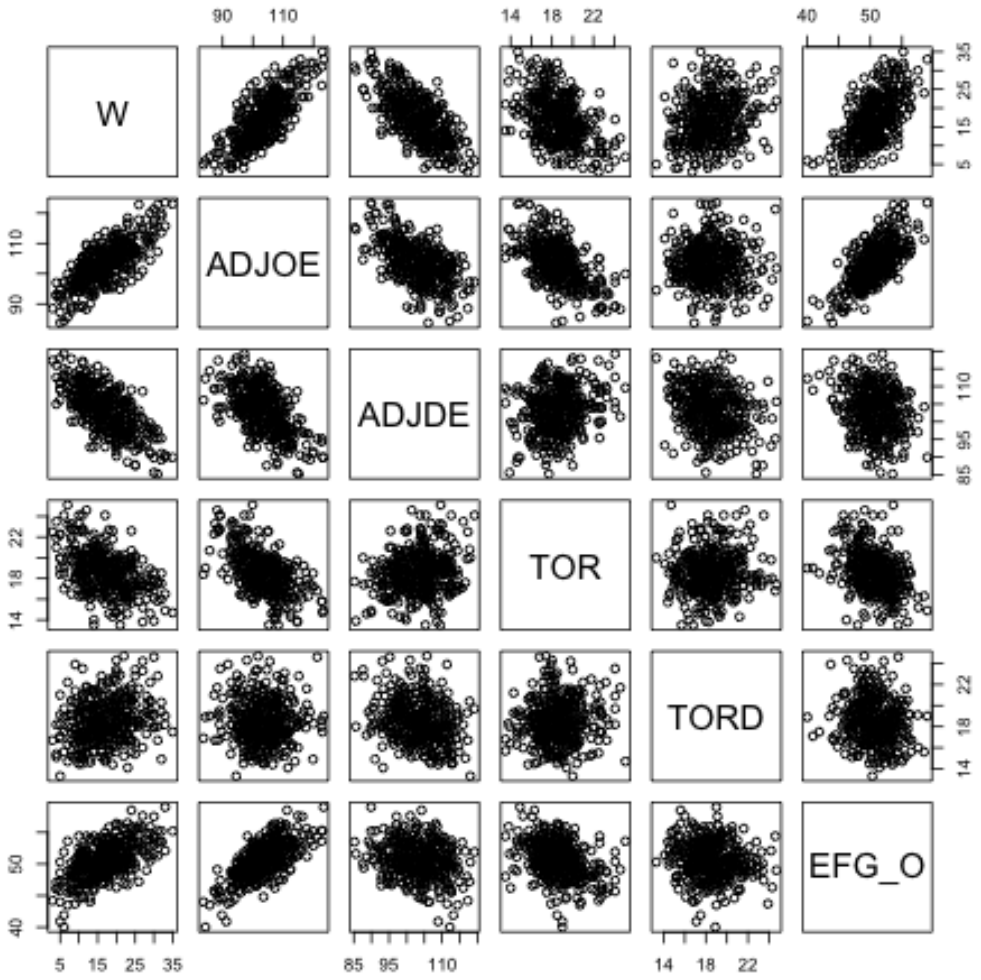






Additionally, I created a correlation matrix along with a graph of some pairwise interactions.

	W	ADJOE	ADJDE	EFG_O	TOR	TORD
W	1	0.74763109907464	-0.684400288942562	0.587549918737607	-0.457309376864682	0.233868446591693
ADJOE	0.74763109907464	1	-0.568743755483125	0.698596323647121	-0.57555088414802	-0.0293797207378335
ADJDE	-0.684400288942562	-0.568743755483125	1	-0.231537861665069	0.224042007578146	-0.263413045164939
EFG_O	0.587549918737607	0.698596323647121	-0.231537861665069	1	-0.359248754684407	-0.123659931781585
TOR	-0.457309376864682	-0.57555088414802	0.224042007578146	-0.359248754684407	1	0.0729274586562059
TORD	0.233868446591693	-0.0293797207378335	-0.263413045164939	-0.123659931781585	0.0729274586562059	1



4 Analysis- Model Building

During model building, I started off using only a couple quantitative predictors that looked like they had the strongest linear relationships from the EDA to predict the number of games won which include both the regular season and post-season. The three quantitative predictors I chose to use were ADJOE, ADJDE, and EFGO.

First Stage Model

Our first stage hypothesized model was:

- $\mathbb{E}(GamesWon) = \beta_0 + \beta_1(ADJOE) + \beta_2(ADJDE) + \beta_3(EFGO)$

After conducting an F-test, we returned an F test statistic of 263.14 with a significant p-value. Meaning, this first model is adequate for predicting the number of games won.

Qualitative Predictors

Next, I added some qualitative predictors by turning the variable, WAB ($wins - above - bubble$) into levels by certain bins. There were a total of 5 bins with equal lengths to be used as a categorical variable.

The bubble refers to the cut-off between making the NCAA March Madness Tournament and not making it.

These bins along with their dummy coding were:

Base level: -23.2:-16:48

X1: x1 = 1 if [-16.48, -9.56), x1 = 0 if not

X2: x2 = 1 if [-9.56, -2.64), x2 = 0 if not

X3: x3 = 1 if [-2.64, 4.28), x3 = 0 if not

X4: x4 = 1 if [4.28, 11.2), x4 = 0 if not

Our second stage hypothesized model with the dummy variables was:

- $\mathbb{E}(GamesWon) = \beta_0 + \beta_1(ADJOE) + \beta_2(ADJDE) + \beta_3(EFGO) + \beta_4(x1) + \beta_5(x2) + \beta_6(x3) + \beta_7(x4)$

Since our F test was significant, our second model with qualitative predictors is adequate for predicting games won.

Adding Interaction

An interaction I thought could be important was WAB*ADJOE since I believed that the more points a team scores, the more likely they may be to make the March Madness tournament.

My third stage hypothesized model included the interaction between WAB and ADJOE which also returned a significant result from a global F-test and thus is adequate for predicting games won.

- $\mathbb{E}(GamesWon) = \beta_0 + \beta_1(ADJOE) + \beta_2(ADJDE) + \beta_3(EFGO) + \beta_4(x1) + \beta_5(x2) + \beta_6(x3) + \beta_7(x4) + \beta_8(ADJOE * x1) + \beta_9(ADJOE * x2) + \beta_{10}(ADJOE * x3) + \beta_{11}(ADJOE * x4)$

Evaluating Interaction

Although the complete model which included interaction terms was significant, it may not neccessarily be better than our second stage model. Thus, I wanted to evaluate if adding the interaction bettween WAB*ADJOE was actually significant or not.

I performed a nested F-test comparing my complete model with my reduced model:

- COMPLETE: $\mathbb{E}(GamesWon) = \beta_0 + \beta_1(ADJOE) + \beta_2(ADJDE) + \beta_3(EFGO) + \beta_4(x1) + \beta_5(x2) + \beta_6(x3) + \beta_7(x4) + \beta_8(ADJOE * x1) + \beta_9(ADJOE * x2) + \beta_{10}(ADJOE * x3) + \beta_{11}(ADJOE * x4)$
- REDUCED: $\mathbb{E}(GamesWon) = \beta_0 + \beta_1(ADJOE) + \beta_2(ADJDE) + \beta_3(EFGO) + \beta_4(x1) + \beta_5(x2) + \beta_6(x3) + \beta_7(x4)$

Our result returned a significant p-value which implies that our parameters are significantly different from 0, and that the interaction between WAB*ADJOE is significant. Our complete model is better than our reduced model. Which leaves us with our final model as the "Complete" model above.

5 Checking Model Assumptions

The assumption that the mean of our error term ϵ held true from our residual plot since we don't see any patterns in the Residuals vs. Fitted graph.

Our scale location plot is approximately horizontal with pretty much even spread around it, which means our assumption of constant variance is met.

To check our normality assumption, we can look at our QQ plot. Our line approximately follows the dashed diagonal line. However, the front and the tail of the data seem to deviate a little.

We can also see that we have several influential observations with a couple of outliers.