MAS 637 - Project

Andrew Wolfe, Julian Pollak, Saul Campanella

Abstract

Linear regression analysis has been one of the most important processes that helps analysts determine and understand the relationship between two or more variables. This project report presents an in-depth analysis of the National Collegiate Athletic Association (NCAA) men's basketball throughout their regular season, analyzing their performance by focusing on significant variables. Keeping the objective of discovering which statistical metrics have the greatest effects in predicting a team's success during a given season, the usage of ANOVA testing, hypothesis testing the Effective Field Goal percentage coefficient, and fitting a parsimonious model with "Wins" as a response variable helped us attain the results of our research. We discovered that Adjusted Offensive Efficiency, Turnover Percentage, Effective Field Goal Percentage, Turnover Percentage Allowed, Free Throw Rate, Adjusted Defensive Efficiency, Three Point Percentage Allowed, and Two Point Percentage Allowed contribute to a team's success during a given season. These findings provide valuable insights that can lay the groundwork for future research in the field of sports analytics, specifically men's collegiate basketball.

Introduction

For our Linear Regression project, we chose the dataset cbb.csv from the internet. This data set has various metrics of how college basketball teams performed in relation to many variables including their record (wins vs losses), Field Goal percentage, seeding in the NCAA tournament, tournament finish, and many other metrics. We intend to fit a parsimonious model using stepwise.

We want to discuss how wins vary by Power 5 conference in the NCAA while doing an ANOVA test through running a linear regression model. We plan to drop certain columns as they are not relevant to our data. These include "G", "SEED", "YEAR", "WAB", "POSTSEASON", "BARTHAG", "TEAM". We then plan to build some scatter plots using the plot and corrplot function to determine correlation between variables, especially variables that affect the number of total wins. We plan to do a histogram of the residuals to determine if this data set has a normal distribution, while simultaneously doing a correlation plot aligning the residuals with the fitted values. This determines whether the values are correlated with the line of best fit. Variables include:

- "G": Games
- "SEED": Seed of teams
- "WAB": Win above bubble
- "POSTSEASON": Postseason result
- "BARTHAG": Estimate of what a team's chance of winning would be against the average DI team
- "TEAM": Team

Our motivation for choosing this data was because all of us in the group are avid college basketball fans, and we thought this detailed data set could impact our learning in the Linear Regression course. We were all curious which statistics were most important in influencing a team or a conference's total number of wins in a season or over a multi-year period. As we do more research and create our plots, we will determine which stats have the highest influence on the total number of wins.

Analysis

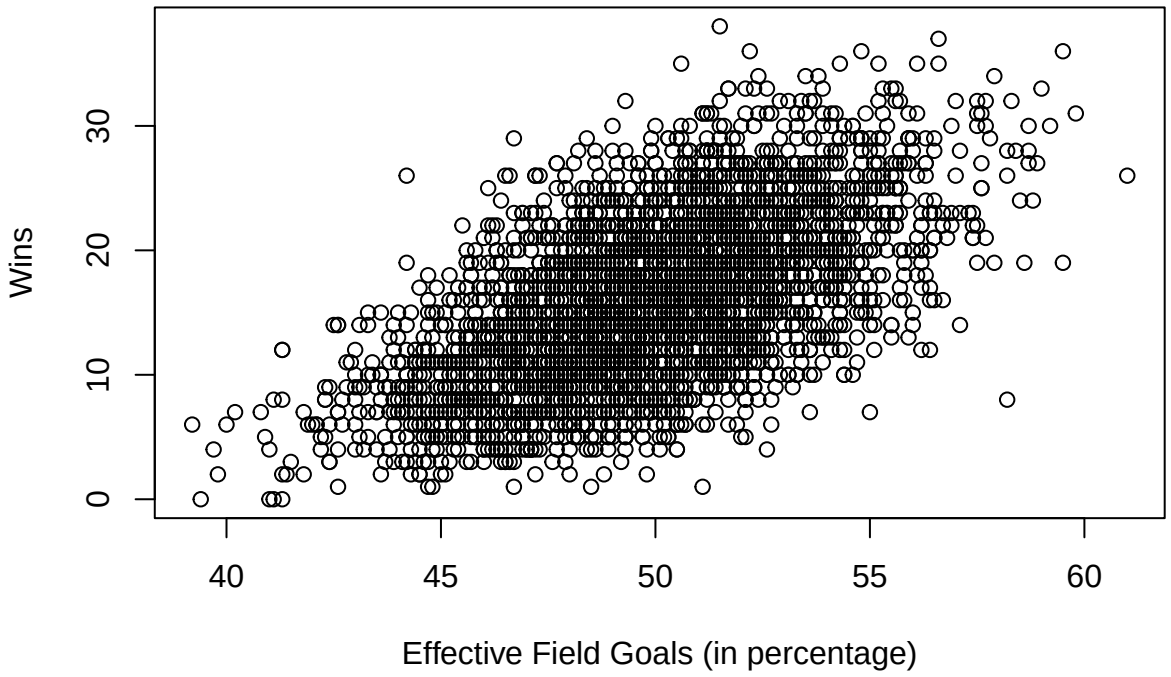
Initially to start our analysis, we conducted an anova test to see the impact that College Basketball Conferences had on winning games. When looking at our anova result, we see that "W", or wins, is the dependent variable and "CONF", or conference, is the independent variable. We are most concerned with the F statistic and p-value, which will determine whether the variations in wins between conferences is statistically significant. Our F-value is 1.2302 and our resulting p-value is 0.2967, which is extremely high and greater than the significance level of 0.05. This indicates that there is not enough evidence to suggest that there are significant differences between conferences in the number of wins.

We also developed a parsimonious model to discover the most important predictors for winning games in college basketball, using general team statistics. To do this, we removed the following columns from our dataset: "G","SEED","YEAR","WAB","POSTSEASON","BARTHAG","TEAM","ORB","DRB", "FTRD", and "CONF","ADJ_T". We picked these variables to be removed because we just wanted to focus on general game-to-game statistics for each team. We use a stepwise regression model and from it, we can see that the regression model formed was:

$$W = ADJOE + TORD + EFG_O + TOR + FTR + ADJDE + X3P_D + X2P_D + \varepsilon$$

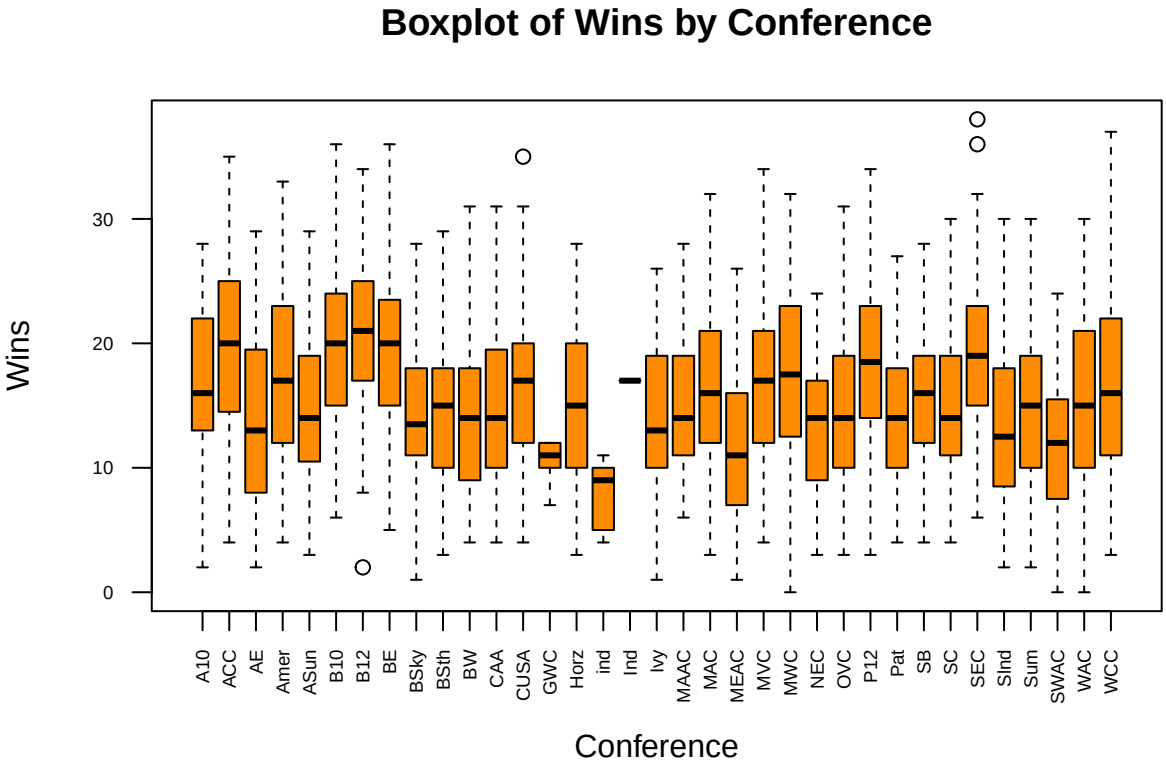
The first thing that stands out is the F statistic and p-value score. The overall p-value is extremely low, indicating that this model is statistically significant and at least one the predictors has an effect on the number of wins for each college basketball team (explanatory variable). Additionaly, it appears that the strongest predictors are "ADJOE," "TORD," "EFG_O," "TOR," "FTR," "X3P_D," and "X2P_D" as their p-values are all equal and they have three asterisks, which suggests very high significance.

Lastly, we created a hypothesis test to see the effect that Effective Field Goal Percentage had on winning. When looking at our 95% Confidence Interval, we see that the predicted number of wins increases between 1.152387 and 1.441741 wins for every one percent increase in Effective Field Goal Percentage. Moreover, the interval does not contain zero, which reveals that the Effective Field Goal Percentage predictor has a statistically significant impact on the number of wins in the model at the 5% significance level.

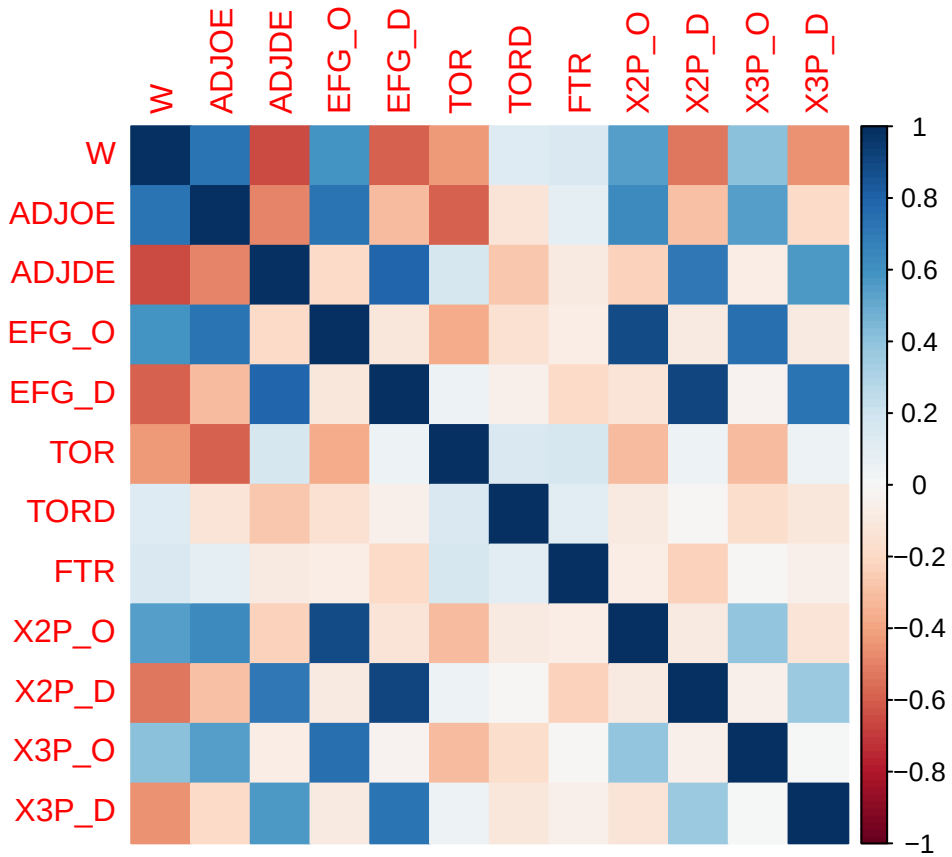


From the plot above, we can see that there is a clear trend/shape with the relationship between Wins and Effective Field Goals for a men's college basketball team. The points are mostly clustered for teams with an effective field goal between 45% and 55%, which in return gave them between 10 and 30 wins over a given regular season.

There are a few outliers, that can be interpreted in two ways depending on their effective field goals. For team with effective field goals lower than 45%, the number of wins are significantly smaller, ranging from 0 to 10. In the other hand, for teams with effective field goals percentages higher than 55%, wins range from 10 up to 30 or more.



By looking at the boxplot above, we see the variations in wins per conference. This graph reveals the large variance and ranges in wins between teams in each conference. There are a few outliers in the SEC and Conference USA, where teams won an exceptionally number of games. One of these points refers to the 2015 one-loss University of Kentucky team. Moreover, the boxplot is significant because it tells the common-reader that wins vary by conference and teams due to many statistics. Our job here is to figure out which statistics are the most important teller for predicting wins.



From the correlation matrix above, we can identify the variables in our data set to examine how these variables are related to one another. With a focus on "W" (Wins), we can see that the following variables have the most positive correlation:

- "ADJOE": Adjusted Offensive Efficiency
- "EFG_O": Effective Field Goal Percentage (Offense)
- "X2P_O": Two-Point Field Goal Percentage (Offense)
- "X3P_O": Three-Point Field Goal Percentage (Offense)

We can also see which variables have the most negative correlation with "W" (Wins):

- "ADJDE": Adjusted Defensive Efficiency
- "EFG_D": Effective Field Goal Percentage (Defense)
- "TOR": Turnover Rate
- "X2P_D": Two-Point Field Goal Percentage (Defense)
- "X3P_D": Three-Point Field Goal Percentage (Defense)

The remaining variables, which are "TORD" (Turnover Rate Offense) and "FTR" (Free Throw Rate) have little to no correlation with "W" (Wins)

Modeling

#Backward Elimination

#Forward Selection

#Stepwise

```
summary(SW)
```

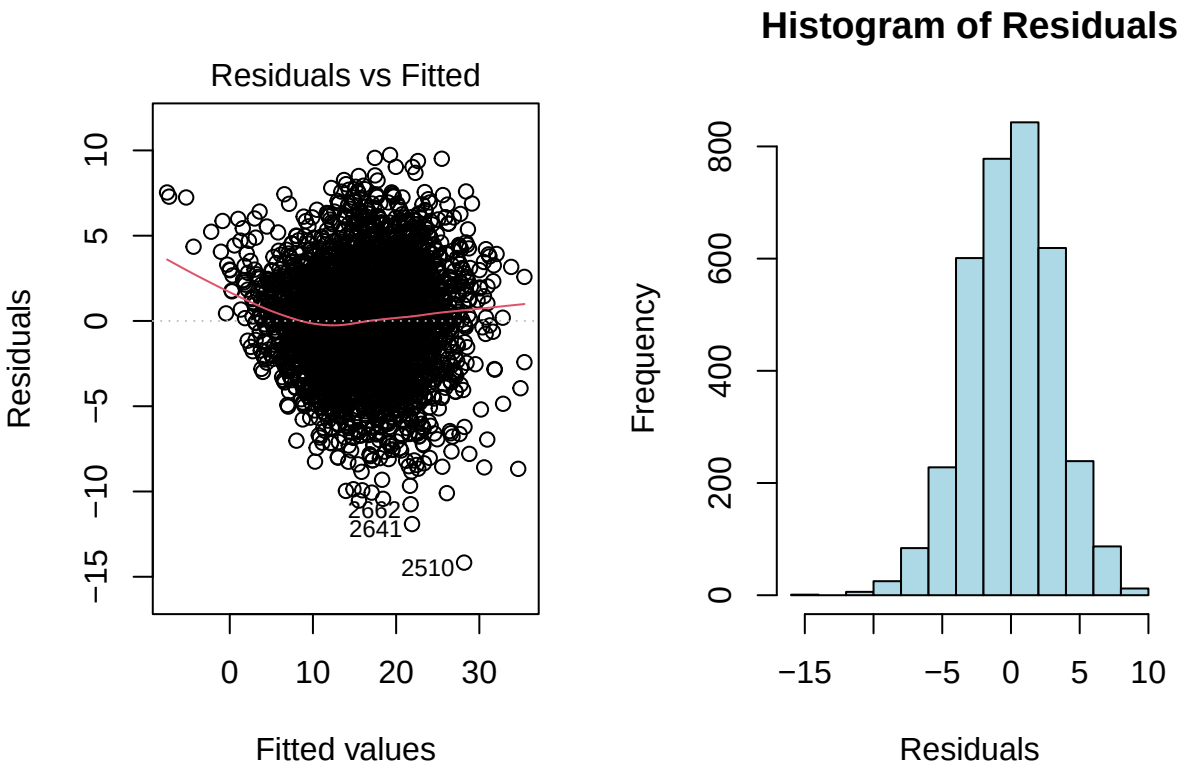
```
##
## Call:
## lm(formula = W ~ ADJOE + TORD + EFG_O + TOR + FTR + ADJDE + X3P_D +
##     X2P_D, data = ncaa_1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.1697  -2.1258   0.0582   2.1693   9.7352
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.05818    2.43328   2.079   0.0377 *
## ADJOE        0.24728    0.01641  15.069  < 2e-16 ***
## TORD         0.54372    0.02816  19.310  < 2e-16 ***
## EFG_O        0.66031    0.02756  23.960  < 2e-16 ***
## TOR         -0.52534    0.03432 -15.308  < 2e-16 ***
## FTR          0.09839    0.01105   8.906  < 2e-16 ***
## ADJDE       -0.08577    0.01787  -4.799 1.66e-06 ***
## X3P_D       -0.51118    0.02762 -18.507  < 2e-16 ***
## X2P_D       -0.50320    0.02557 -19.679  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.168 on 3514 degrees of freedom
## Multiple R-squared:  0.7683, Adjusted R-squared:  0.7677
## F-statistic:  1456 on 8 and 3514 DF,  p-value: < 2.2e-16
```

For our parsimonious model, we used stepwise regression. Once testing out forward selection, backward elimination, and stepwise regression, we noticed that our best model was with stepwise regression. Stepwise regressions start with no independent variables and adds variables one at a time if SSE can be significantly reduced. In this case, our stepwise regression model gave us an adjusted r-squared of 0.824. We were attempting to discover which variables impacted winning the most in college basketball during the last 10 seasons.

Also, from the above results, we discovered that the Backward Elimination and Stepwise had the same BIC at 18194.65, while the Forward Selection was different at 18201.18. When comparing the BIC of each parsimonious model, we saw that the BIC for stepwise and backward elimination was lower than forward selection. Stepwise considers eliminating some of the variables, which is something that forward selection does not do.

Model Assumptions

The assumptions are pretty well met. The linear regression assumptions are given by: LINE. For the non-linear component, there appears to be a slight negative slope at the beginning, but it is later stabilized over the line of best fit.



By looking at our parsimonious model, we see that Adjusted Offensive Efficiency, Turnover Percentage, Effective Field Goal Percentage, Turnover Percentage Allowed, Free Throw Rate, Adjusted Defensive Efficiency, Three Point Percentage Allowed, and Two Point Percentage Allowed are the 8 most valuable predictors our model for statistics that contribute to winning the most.

Effective Field Goal Percentage Analysis

Conference Analysis

Null Hypothesis: Wins are the same across the 5 different conferences for men's college basketball teams.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

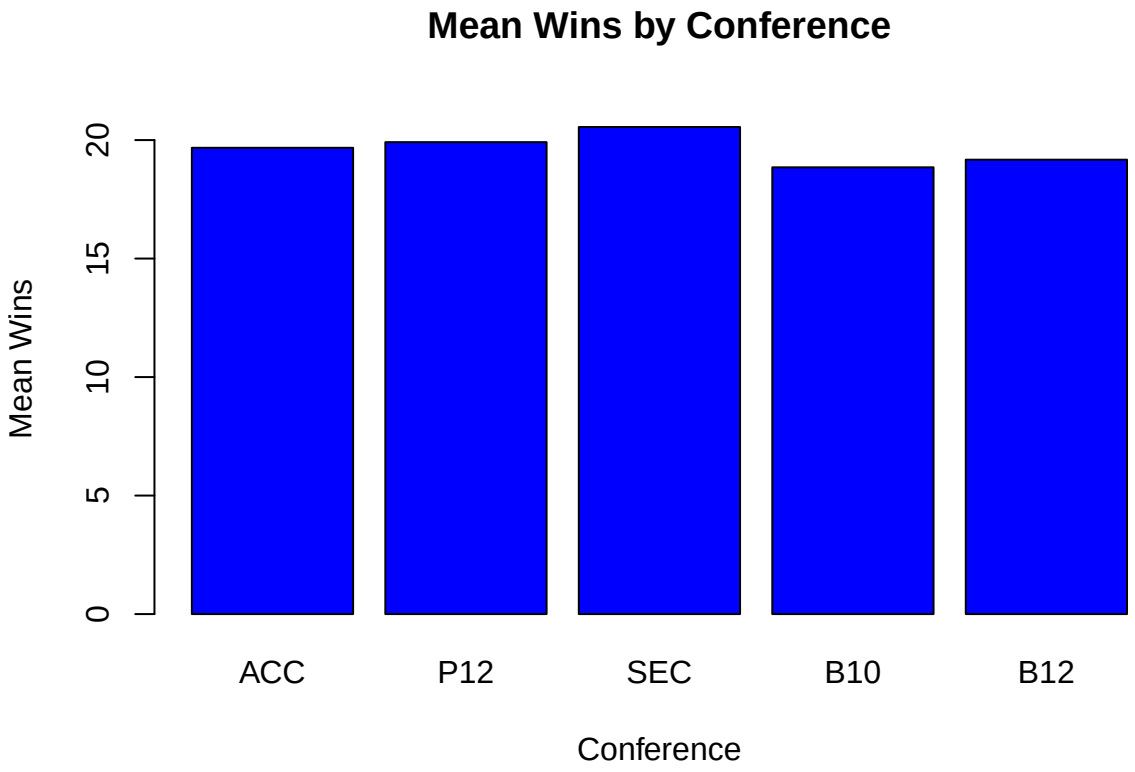
Alternative Hypothesis: At least one conference will have a different "Win" amount.

```
##
## Call:
## lm(formula = W ~ EFG_O, data = ncaa5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.7436  -3.6411  -0.0407   3.3945  17.2732
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -46.07220    3.73571  -12.33   <2e-16 ***
## EFG_O         1.29707    0.07367   17.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.209 on 641 degrees of freedom
## Multiple R-squared:  0.3259, Adjusted R-squared:  0.3249
## F-statistic:   310 on 1 and 641 DF,  p-value: < 2.2e-16
```

From the results, taking into consideration the null hypothesis, which states that wins are the same across the 5 different conferences for men's college basketball teams, and also the alternative hypothesis, which states At least one conference will have a different "Win" amount, we conclude that the alternative hypothesis is correct. Examining the results, we can see that even though the amount of wins during a given season across the 5 different conferences were very close to being the same, there are still conferences that managed to average more wins than others. We reject the null hypothesis.

Bullet 1 (ANOVA Test)

```
## Analysis of Variance Table
##
## Response: W
##            Df  Sum Sq Mean Sq F value Pr(>F)
## CONF        4   197.5  49.368  1.2302 0.2967
## Residuals 638 25603.2  40.130
```



From the plot above, we can see that the mean wins per the top 5 conferences in NCAA men's basketball is very close; they all fall between the range of 17 to 20 wins. This shows how competitive the Top 5 conferences are, and proves to show that factors, such as effective field goal percentage, can be fundamental for the success of a team under such a competitive tournament.

Bullet 2 (Confidence Interval)

Null Hypothesis: A coach claims that having a high team effective field goal percentage (EFG%) increases wins in an overall season.

Alternative Hypothesis: A high team effective field goal percentage (EFG%) does not increase wins in an overall season.

$$1.15 \leq \beta_{EFG} \leq 1.44$$

Based on setting up a hypothesis test on the EFG% coefficient, the model predicts that the beta coefficient for EFG% should fall within the range of 1.15 to 1.44 at a 95% confidence.

From the results, taking into consideration the null hypothesis, which states that coach claims that having a high team effective field goal percentage (EFG%) increases wins in an overall season, and also the alternative hypothesis, which states high team effective field goal percentage (EFG%) does not increase wins in an overall season, we conclude that the null hypothesis is correct. With such a competitive tournament like NCAA men's basketball, there are key variables like effective field goals that affect a team's performance throughout a season. Teams with higher effective field goal percentage ended up with a higher number of wins as opposed to other teams with a smaller effective field goal percentage. We accept the null hypothesis.

Conclusion

All in all, we determined the null hypothesis was correct that the Adjusted Offensive Efficiency, Turnover Percentage, Effective Field Goal Percentage, Turnover Percentage Allowed, Free Throw Rate, Adjusted Defensive Efficiency, Three Point Percentage Allowed, and Two Point Percentage Allowed are the 8 most valuable predictors our model for statistics that contribute to winning the most. There was strong correlation between our predictor variables and the total number of wins. After we completed our ANOVA test and stated out null and alternate hypothesis, we decided that our predictor variables definitely affect total wins amongst the Power 5 conferences that we studied. This showed that we accept the null hypothesis and reject the alternative hypothesis. Throughout this process, we learned the importance of linear regression analysis to determine how predictor variables affect an outcome. This dataset was interesting to us because we are all avid college basketball fans and enjoy learning about the statistics of the game.

This conclusion is extremely significant because coaches can use it to help in their player development, in-game coaching adjustments, and scouting analytics. For instance, by looking at the model, we that forcing turnovers from your opponents, scoring at an effective rate combination (with twos and three point attempts), and the ability of the team to defend the two and three point shot are all significant factors in helping your team. All of these predictors can be easily practiced and game-planned for before games. Coaches can use this report to put stronger emphasis on these statistics to their players and in practices.

Also, in the realm of the transfer portal (College Basketball Free Agency), this linear model and 8 valuable predictors can be crucial. For instance, coaches can target players who excel at these statistics. Certain players might have a higher Effective Field Goal Percentage, lower Turnover Percentage, etc., but coaches can tailor their team needs to players who are specifically strong in certain statistics that have been proven to great correlate with winning. In conclusion, this model and project will be critical for college basketball coaches are trying to improve their roster for the future and at the same time, in the present with scouting and player development.