John Keenan, Andy Beckham, Andrew Lege

**Introduction:**

FBS college football is unique. Millions of Americans attach their happiness to the successes and failures of college students playing a football game. This passion extends beyond the field, where TV networks, beat reporters, and recruiting analysts rush to cover games, break news, and offer different insights far beyond the few fall Saturdays when games take place. In all this analysis, from breaking down defensive coverages to rating high school prospects, one goal trumps all: winning. However, this analysis frequently rushes to the particular and ends up overlooking the broad examination of the characteristics of winning teams.

Data Analysis in college football is scarce. Outlets such as Pro Football Focus and SportsDataIO publish data analysis with an emphasis on the NFL draft, fantasy football, and sports betting. However, their research focuses on individual players as opposed to team statistics and is behind a paywall, making it not widely accessible. College Football Reference provides data about teams and players, but it only presents data without any analysis, making it difficult to see which statistics are most important. Unlike the NFL and other professional leagues, there is no player tracking technology, adding to the difficulty of drawing conclusions from the data.

In 2019, 130 teams played Division I football. Data was recorded on these teams with many classes and variables present. Each of these 130 teams make unique decisions based on many key factors, including player talent, conference, and coach specialties. These decisions change the outcome of every play, but how do these decisions and outcomes directly affect wins and losses? The goal of this paper is to answer this question by analyzing 2019 Division I data from the NCAA on offensive, defensive, and special teams decisions and outcomes in order to discover the traits of both winning and losing college football teams.

**Methodology:**

The data for this regression analysis was taken from the National Collegiate Athletic Association (NCAA) website, specifically using their dataset from the 2019-2020 Division I college football season.[[1]](#footnote-0) During each week of the college football season, beginning in August and ending in January, the NCAA tracks team performance in a variety of offensive, defensive, and special teams categories, as well as overall record for each team. In the dataset there is a single row containing 127 variables for each team. For instance, each team has a statistic for the variables offensive yards, defensive plays, and average yards per kickoff return.

This data was collected over both the regular season and the postseason (bowl games and the College Football Playoffs). These statistics are updated weekly, enabling users to gain information on team play calling tendencies, strengths, and weaknesses. Each data point is carefully checked and verified using in-game footage, and is updated in the case of any correction being needed. Data is tracked live each game and the statistics are updated in the season-long dataset at the end of the game. By the end of the year, after 12 to 15 games (depending on the success of the team) have been played, all data is inputted.

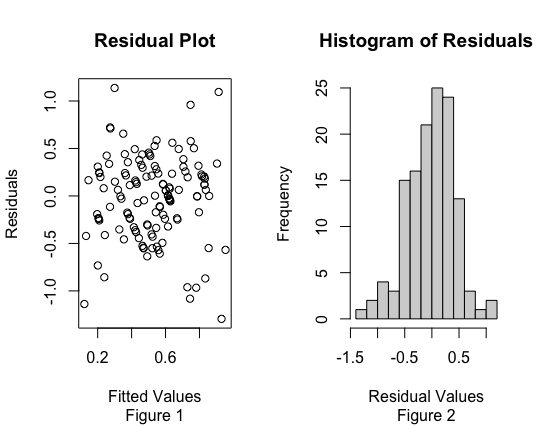
The goal of this analysis is to understand which variables contribute to winning games. In the analysis, win percentage was the only response variable used. However, not all 127 variables are used in the analysis. Our analysis focuses exclusively on variables that were deemed to be independent of the win probability at which the plays that come to comprise the final statistic occur.

**Analysis:**

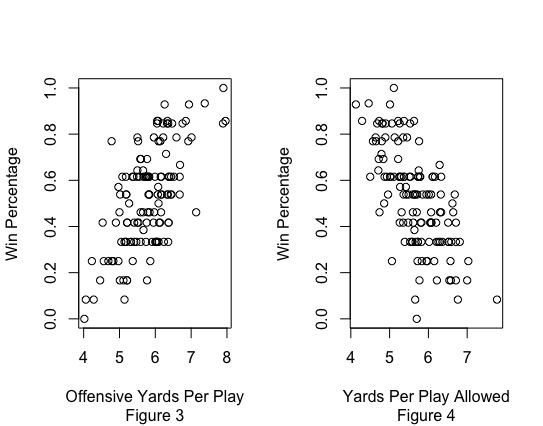
A model was created predicting win percentage from all variables deemed to be independent of game context using binomial logistic regression. Since win percentage must be between 0 and 1, logistic regression was used. Furthermore, since teams play different numbers of games, as the best teams play conference championship and bowl games, binomial logistic regression was used. Each of the 128 variables in the complete data set was determined to be either dependent or independent of game context, and context independent variables were included in the model. For example, rushing attempts is not independent of context, as a team that is leading will run more to avoid clock stoppages. On the other hand, passing yards per attempt is independent of context since it is an average and not especially dependent on game script over the course of an entire season. Additionally, selected interactions based on knowledge of the sport of college football were included in the original model.

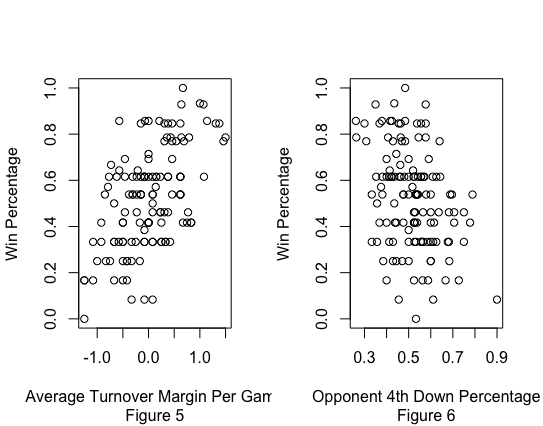
The following 21 explanatory variables chosen in the original model were all determined to be independent of context: offensive yards per play, yards per play allowed, average turnover margin per game, average time of possession, fourth down conversion percent, opponent fourth down conversion percent, average yards per kickoff return allowed, average yards per kickoff return, pass yards per attempt, pass yards per attempt allowed, penalty yards per game, average yards per punt return, average yards per punt return allowed, yards per rush allowed, yards per rush, average sacks per game, third down conversion percent, fumbles recovered per game, fumbles lost per game, interceptions thrown per game, and defensive interceptions per game. The interactions between offensive yards per play and pass yards per attempt, offensive yards per play and yards per rush, pass yards per attempt and yards per rush, yards per play allowed and pass yards per attempt allowed, yards per play allowed and yards per rush allowed, as well as pass yards per attempt allowed and yards per rush allowed were also included.

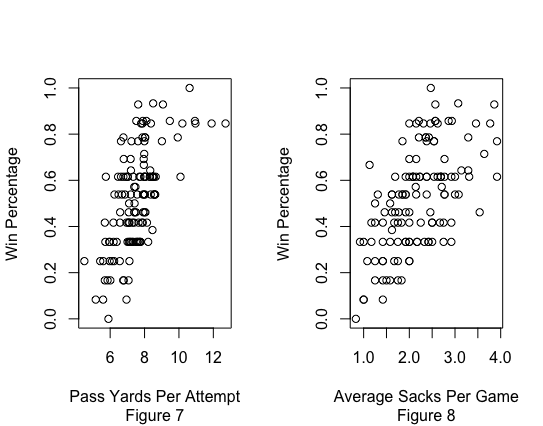
From this initial model, Akaike Information Criterion (AIC) was used to perform criterion-based variable selection. Through backward selection, we removed predictors one at a time until the AIC value was minimized, giving us the optimal subset of predictors. After undergoing this process, we were left with a reduced model using the following predictors: offensive yards per play, yards per play allowed, average turnover margin per game, opponent fourth down conversion percent, pass yards per attempt, average sacks per game, third down conversion percent, fumbles recovered per game, fumbles lost per game, interceptions thrown per game, defensive interceptions per game, yards per rush allowed and the interaction between yards per play allowed and yards per rush allowed.

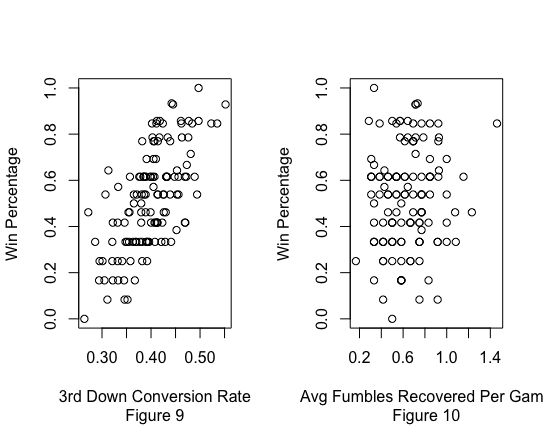


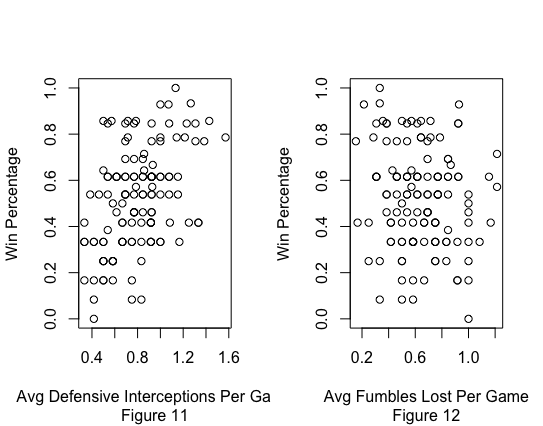
This reduced model, which forms the basis of our analysis for the remainder of the paper, meets the assumptions for model inference. Based on figure 1, the average of the residuals is zero. There is no pattern in the residual plot, so the residuals have constant variance and are independent and identically distributed around zero. Based on figure 2, it can be concluded that the residuals are approximately normally distributed as the curve is approximately bell shaped. Based on figures 3-14, all of the variables in the model are linearly related to win percentage. There is no need to include quadratic or higher order terms.

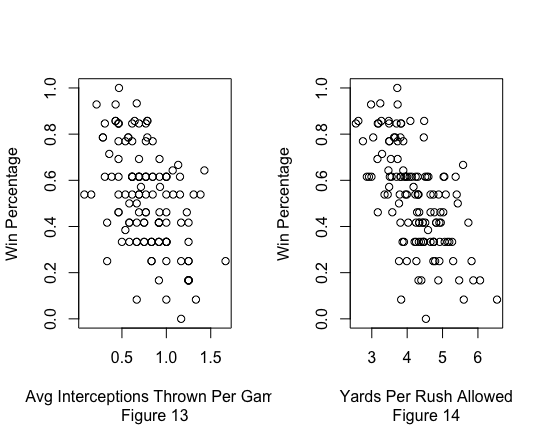
****





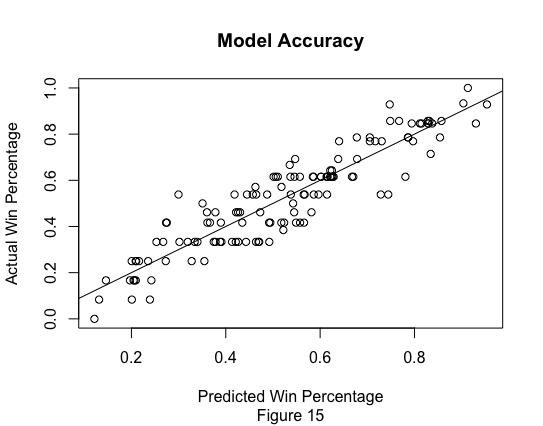






Overall, this reduced model provides a strong fit in predicting win percentage for NCAA Division I football teams. It fits the four assumptions of a linear regression model and, based on our initial selection of predictors, should eliminate any variance due to game script trends. We will continue to use the reduced model for the rest of our discussion for these reasons.

**Discussion:**

****

As seen in figure 15, the model effectively predicts winning percentage. It captures the characteristics of winning teams. Winning teams maximize yards per play and limit their opponents’ yardage. Takeaways, whether fumbles or interceptions, promote winning, but turnovers, both fumbles and interceptions, promote losing. Teams should strive to convert on third down while preventing opposing teams from converting on fourth down when they do not punt. All of the model’s traits associated with winning reaffirm the common wisdom of fans and coaches.

Although none of the characteristics of winning teams are surprising, the model’s insight into the relative value of certain traits is striking. For example, fumbles and interceptions are of equal value in predicting winning. A forced fumble has the same slope coefficient as a defensive interception. That slope coefficient is the opposite of the slope coefficient of an offensive fumble or interception and average turnover margin per game. The model also provides insight into the relative value of rushing and passing. In variable selection, offensive yards per rush and defensive yards allowed per pass attempt were deemed insignificant. While gaining yards on offense and limiting them on defense is most important, winning teams are more likely to have a thriving passing offense and a defense that stops the run.

Despite offering these important conclusions, this analysis has clear limitations. Sacks given up, opponent penalty yards per game, and opponent 3rd down percentage are potentially important statistics that are not reported in the data. More broadly, this analysis is focused on an entire season, not single games. As such, conclusions about in game decision making, specifically clock management and play calling, cannot be drawn. Most glaringly, this analysis is based on the assumption that the variables input into the model are context independent. Without game data, or even more specific play logs, it is impossible to test this assumption.

This research into the characteristics of winning college football teams, even with its clear limitations, can be of use to future investigators of the sport. First, this analysis reveals that better data must be collected on individual games, players, and plays to more thoroughly explore the most important characteristics of winning teams. Second, the traits of winning teams identified by this model can be used in more granular analysis. For example, yards per play could be used as a measure of success in analyzing offensive personnel groupings or play sequencing. Third, the model prompts future investigators to pursue truly context independent statistics that improve future analysis. Overall, this model provides a solid foundation for future analysis by confirming many of the assumptions of college football fans, coaches, and analysts. However, further data is required to create a more useful model relating in-game decisions to success on the field.

**Conclusion:**

In a college football environment lacking public data analysis, this research is a first step to better understanding the sport. This analysis of season statistics for each of every team in FBS college football attempts to determine which context independent statistics best predict success on the field: winning football games. Although the resulting model does yield any surprising conclusions, it affirms the assumptions of the TV networks, beat reporters, and recruiting analysts who dominate the discussion of the sport. Moreover, the confirmation of the value of gaining yards on offense, limiting yards on defense, and getting takeaways while limiting turnovers provides a foundation for future research of individual game and individual play data.

1. All data comes from the NCAA’s official team statistics found at <https://www.ncaa.com/stats/football/fbs>. [↑](#footnote-ref-0)