Predicting College Football Wins

By: Jordan Clarke

Abstract:

The idea behind this paper was to develop an accurate and enduring formula for predicting the success of college football teams throughout the course of the college football season. Contrary to its NFL counterpart, College football is extremely difficult sometimes impossible to predict at the beginning of a season. For example Alabama went undefeated almost two seasons in a row until Clemson defeated them in the National Championship game. Given our interest in college football, we wanted to see if we could find a statistical model that could predict the success of college football teams more effectively than my peers. We will predict the number of wins in a season for a college football team using a regression. This model will include statistics like: pass yards, rushing yards, rushing defense, passing defense, and penalty yards. These values will all be in yards per game. The goal for this paper will be to develop this model for the Power 5 conferences to have enough data to lead to significant values. After the regression is run we will see which values had a p value of less than .05 in order to make our final model. We will then test this model by inputting previous season's statistics and compare the model prediction to the actual wins by each team. We will then use the model to try and predict future wins, although this will be difficult to achieve accurately because of the many departures in the world of college football. We will use each team's statistics from the ESPN website and divide the season totals by the number of games each team played. When examining the best of the best in college football, it's all about the Football Bowl Subdivision also known as the FBS.

Table of Contents

I. Introduction:

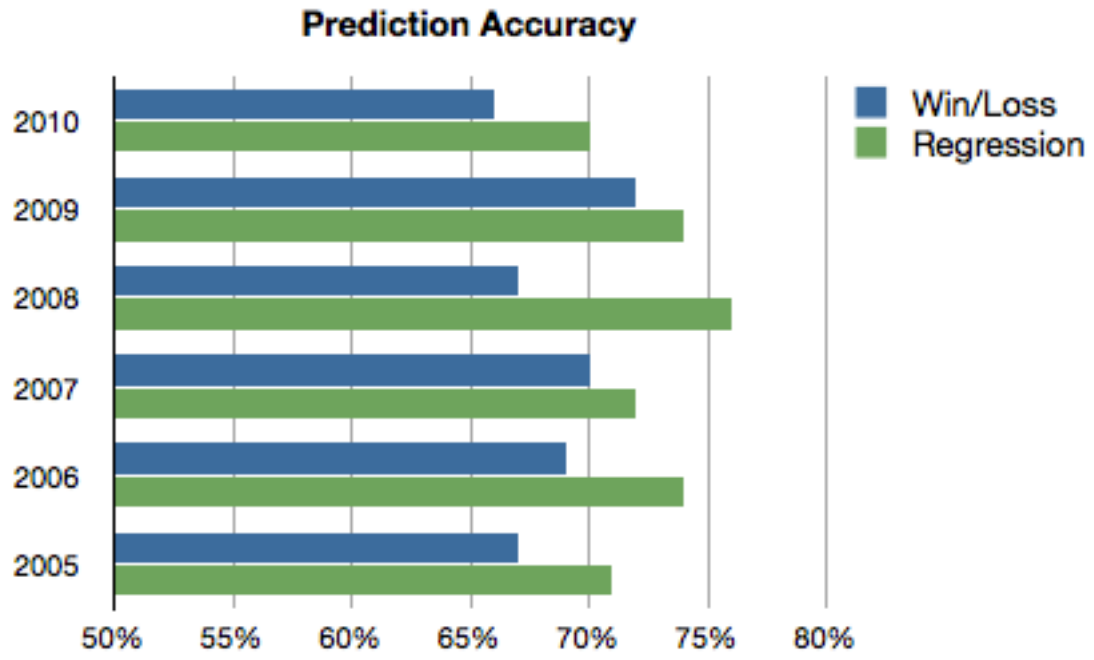
It's the beginning of the New Year, The Clemson Tigers are playing the Alabama Crimson Tide once again and you're watching the game in the family room with nobody else but family. This wonderful sight is the scene in many Americans households. College football is among the most popular sports in the United States. There are some powerhouse college football teams. Some states breed star studded football players who grew up dedicated to their respective state universities and other schools have incredible recruiting powers. Then, there are schools that bring in star-studded coaches that bring their program up to the next level. Time and time again we see these powerhouses win national championships. We see exceptional coaches move from university to university and create winning teams. At the beginning of the college football season all teams are hungry for the title of National Champion, however many teams do not get the opportunity to be considered for the National Championship.

The History of college football, a spectator sport in the United States, can be traced to early versions of rugby (Watterson). Both games origin originated from football played in the United Kingdom in the mid-19th century, in which a football is kicked at a goal and run over a line (Watterson). The popularity of college football grew, as it became the dominant version of football in the United States. Bowl games attracted a national audience for collegiate teams. "With a lot of fierce rivalries, College football still holds the widespread popular appeal in the U.S. Intercollegiate athletics in the United States has come to be regarded as higher education's "peculiar institution." This critical characterization results from the fact that college sports are listed as part of the central mission of colleges and universities. Visitors to American campuses cannot help but to be struck by the presence of college football teams. In the twenty-first century,

it is not unusual for major universities to contain both a football stadium that seats over 50,000 spectators. In the early 2000s, many universities operating budgets for athletics ranged between 30 and 60 million.

The average revenue to each of the power five conferences in 2014-15 was around $70 million from the playoff system compared to $30 million before the playoff system (Ridpath). Each of the 10 FBS conferences receive a base amount combined with an academic performance pool is approximately $55 million for each conference (Ridpath). Conference that reaches the semifinal (Power 5 Conferences) receive a bonus $6 million (Ridpath). Each team that makes it to the national championship game will receive an additional $2.16 million to cover expenses (Ridpath). This money normally goes toward the Power 5 conferences since teams out of these conferences normally make it to the playoffs. The Playoffs consist of eight teams with the number increasing year per year with the value increase to more than a billion per year. The NCAA revenues and majority of all college football playoff revenues go to the Power Five conferences, which reveals the dominance of division 1 football programs. The goal of the Power Five conference institutions is to win and acquire the greatest amount of revenues possible in the process of doing so. Most will spend whatever it takes to building a successful program so that they can control the distribution of revenues. Even though college football programs have an enormous stream of revenues it still doesn't guarantee wins.

The use of statistics is an integral component of sports. The number of touchdowns scored in football is means by which analyst and fans alike determine success for specific players and teams. The purpose of this paper is to present a regression model for analyzing game statistic important in determining the outcome of College Football Games.



II. The Model

The model that we created in order to estimate college football wins considers many of the variable we think contribute to scoring the most points, and the opposing team from scoring as little points as possible. The variables we used to estimate this are: rushing yards per game, passing yards per game, opponent rush yards per game, opponent pass yards per game, turnover margin, and field goal percentage. The variable turnover margin was included with these other traditional point scoring variables in order to account for teams that may do well statistically, but are prone to make mistakes throughout a game. The intercept variable is the wins each team had in the 2016 regular season.

Rushing Yards

This variable was included because of the importance of running the football in order to protect the quarterback and control the tempo of the game. If a team can run the football well in a game, they can control the clock and will take over the game once the opposing defense starts to give in. This becomes very apparent in the later part of a game, especially the fourth quarter. This can be seen with teams like Alabama which had the second highest in the Power Five Conferences with an average of 250.6 yards per game. This helped Alabama have the highest number of regular season wins at 12. This is also true at the other end of the scale as teams at the bottom like Purdue, which averaged only 81.4 yards rushing per game are kept from maintaining any control over the game and was a factor that led them to only winning 3 games.

Passing Yards

The next variable we look at is passing yards. This can be very important to a team because it is usually the quickest way to get points which becomes very important at the end of each half. This ability to quickly score points can help a team tremendously when the game is close and comes down to the last few minutes. The importance of this can be seen in a team like Clemson, who wasn't one of the higher rushing teams, but had the fourth highest passing yards per game at 334.4. This helped Clemson win many close games and come out of the regular season with 11 wins. The opposite also holds true for this as a team that can't effectively pass the ball has little chance to run the ball either since the defense will know exactly what is coming. This can be seen in a team like Rutgers that averaged only 138.4 yards passing per game and only came out of the regular season with 2 wins.

Opponent Rushing Yards

This is a very important variable for the same reason rushing yards is a very important variable for the offense. If your defense can keep a team from running the ball effectively, it will give them little to no control over the game and will help the defense stay off the field and perform better later in the game. Alabama also did very well in this category and had the lowest number of rushing yards given up per game at 63.9. This was another major factor that helped Alabama win all 12 games in their 2016 regular season. Rutgers was also underperformed in this category and had the highest rushing yards given up at 274.5 yards per game. This kept them from controlling any part of the game helped lead them down to only 2 wins in the regular season.

Opponent Passing Yards

This variable is important in predicting the number of wins for the same reason as the variable of offensive passing yards is important. Anytime your defense can help stop the passing attack it will prevent the opposing team from scoring quickly to get back in the game and keep their offense one dimensional. This can be seen in a team like Michigan which gave up the fewest number of passing yards per game at 142.5. This helped propel Michigan to a 10-win season. The opposite side of this would be a team like Syracuse which gave up a much higher 291.8 yards passing per game. This is a major reason Syracuse was only able to win 4 games this year as they couldn't keep many teams from jumping out ahead of them.

Turnover Margin

This is a very important variable in estimating wins because a team can have more yards on offense and give up the fewest yards on defense, but if they make mistakes and turn the ball

over, especially in their own red zones, they will not be able to win games. A team like Washington did very well in this category with a positive 1.2 turnover margin per game. This helped Washington to 11 regular season wins and a berth in the College Football Playoff for the first time. Purdue is another example of how this variable can affect your chances of winning games as they had the lowest of the Power Five Conferences with a -1.8 turnover margin per game. This is also a reason they won only 3 games in the regular season.

Field Goal Percentage

The last variable we included in our model was field goal percentage. This is a very important variable because if your offense isn't able to get the punch it in and get the touchdown you at least need to have those 3 points almost guaranteed. This is also a major factor in the final minutes of a game because the better your kicker is, the farther you can kick a game winning field goal, meaning your offense doesn't need as much time to get into a conceivable field position to win the game. A team that had a very good 91.67% field goal percentage was Penn State. This helped them to their best season since their sanctions began and helped them get 10 regular season wins.

Error Term

The error term in this model will include everything else that affects how many wins a team gets in the regular season. This includes many things such as: injuries, coaching changes, schedule difficulty, and player motivation. These are just a few examples as there as an inordinate number of things that can affect a college aged player's ability to focus on a game and win as many games as possible for his team.

The Model

Wins = 5.0379 + 0.0225 * Rush_Yards - 0.0246 * Opp_Rush_Yards + 0.0162 * Pass_Yards - 0.0166 * Opp_Pass_Yards + 0.9004 * Turnover_Margin + 2.6559 * FG% + ε

This is the model to come up the estimated number of regular season wins for the 2016 season based on rushing yards, passing yards, opponent rushing yards, opponent passing yards, turnover margin per game, and field goal percentage. The ideal model would be able to predict the regular season wins for the following season, however this would be very hard to implement especially in college football without a way to evaluate the new players each team will have that haven't seen the field before.

III. The Data

The multiple regression consisting of the 64 Power Five Conference teams was run with the full summary statistics in the annex. In this section, we will highlight the most important figures that give us the insight we were looking for to predict the number of regular season wins. First, we will look at the mean for each category to give an idea of what a typical team would put out in each category and compare that value to the 2016 National Champions, Clemson. For variable rush yards per game, the mean was 179.3375, compared to Clemson's 165.9. For opponent rush yards the mean was 174.9969, compared to 135.4 yards per game for Clemson. The next variable, pass yards, had a mean of 239.4125 which is considerably less than Clemson's 334.4. Opponent pass yards had a mean of 232.5094 compared to Clemson's value of 191. Turnover margin had a mean of 0.0859, considerably greater than Clemson's -.2 turnover margin per game and field goal percentage had a mean of .7297 compared to Clemson's .7647. We will

now take a closer look at each of these variables to see what sort of an impact each of them had on predicting the number of wins.

Rush Yards

The variable rush yards had a very high T-Statistic of 5.297 and a very low p-value of 0.0000 meaning the null hypothesis can be rejected up to the 1% significance level. These values tell us that team rushing yards per game do have a significant impact on the number of games a team wins in the regular season.

Opponent Rush Yards

The variable opponent rush yards had the most significant T-Statistic of all the variables at -5.491 with a p-value of 0.0000. This means this variable is also able to reject the null hypothesis up to the 1% significance level. This value indicates that as would be expected, the better your defense is at preventing the opposing team from getting a running game, the more wins they can expect to have in a regular season.

Pass Yards

The variable pass yards had a T-Statistic of 4.975 and a p-value of 0.0000. This means the null is also rejected for this variable up to the 1% significance level, meaning it has a significant impact on predicting the number of wins a team can expect in the regular season.

Opponent Pass Yards

Opponent pass yards had a T-Statistic of -3.444 and a p-value of 0.0011. This p-value also indicates that opponent pass yards is significant and will reject the null hypothesis up to the 1% significance level. This also confirms what most people would think, as pass defense is a very important category and can be the difference in many games.

Turnover Margin

Turnover margin had a lower T-Statistic of 2.465 and a higher p-value of 0.0167. This is less significant than the other variables, but still allows the null hypothesis to be rejected at a significance level of 5%, meaning turnover margin still has a significant impact on the number of games a team wins in a season.

Field Goal Percentage

Field goal percentage was the weakest variable with a T-Statistic of 1.860 and a p-value of 0.0680. This means that field goal percentage isn't able to reject the null hypothesis and isn't significant at the 5% level, however we still think it is a part of winning games that should be included in the model because of how impactful a good kicker can be on winning games.

Implications

The model that the regression comes up with tells us a lot about what is significant in winning as many regular season games as possible. The model tells us that each additional rush yard per game your offense is able to get increases your regular season wins by 0.0225. While this doesn't seem very significant, if you look at it in the case of an attainable goal for a coaching staff, like an additional 30 yards per game, it can increase your expected wins by .675 wins, holding everything else constant. The same can be said for the defense, if you are able to hold the opposing team to just 1 less rushing yard per game, you can expect your wins to increase by

0.0246 wins. Again, this doesn't seem like much, but this is just based off 1 less yard, if you do the same 30-yard goal for the defense it increases the expected wins by .738 wins, holding everything else constant. You can combine these rushing categories together and conclude that if you can get your offense to increase their rushing yards per game by 30 yards, while getting your defense to give up 30 less rushing yards per game, the expected wins in the regular season increase by 1.413, which is very significant with the competitive nature of college football.

We can also use the model to predict the effect of improving in passing offense or passing defense to increase expected wins. Each additional passing yard per game earns a team 0.0162 more expected wins in a season. This again doesn't sound very significant but if you look at the number of what a team might set as a goal for the next season it makes a large difference. If a team could improve their passing yards per game by 50 yards the next season, this would increase their expected wins by .81 wins. On the defensive side, holding your opponent to 1 less passing yard per game increases expected wins by 0.0166 wins. As on the offensive side, if a team is able to decrease their opponent's passing yards per game by 50 yards they will increase their expected wins by .83 wins. If a team could make both of these improvements, they would increase their expected win total by 1.64 wins. This is a major difference and could be the difference between a successful season and a disappointing season.

Looking at these figures it is very easy to see why schools can be so harsh in the firing of their coaches. Because the business of college football is so massive, schools want to be able to make as much money as possible. In order to do this, teams must win as many games as possible and using this model it is easy to see that small improvements can make a big difference in the number of expected wins in the regular season.

Model Performance

In this section, we will look at specific cases of how well the model predicted the number of regular season wins for certain teams. We will look at Clemson first. The model predicted 9.54 wins in the regular season, 1.46 wins off the 11 regular season wins Clemson had. There could be many reasons for this, but I think the main reason was that Clemson was involved in many very close games that went down to the wire, meaning they didn't outperform a lot of teams by very much. Another reason the model has a lower number of wins is that Clemson had a relatively low number of rush yards per game when compared to teams of similar caliber. In the case of Florida State, the model predicted exactly 9 wins, exactly the same number of wins Florida State had in the regular season. Looking at the data, Florida State was very consistent in all of the key statistics in the model, making the prediction very accurate for the expected number of regular season wins. The key for the model making an accurate prediction is being able to measure what it is that makes a team have more wins than another. This can be very difficult for a team like Clemson this year who had a player like Deshaun Watson who could make many mistakes during a game and not have great statistics, but in the final two minutes of the game, will be perfect and lead his team to victory. This is what makes the sport so fascinating, the predictions can be so far off from what actually happens and it can be impossible to know which team will show up to play. This is one of the main reasons that models can be useful with all the data and predict wins in hindsight, but almost impossible to predict wins in the future.

IV. Conclusion

The findings presented in this analysis pertain strictly to college football teams apart of the Power 5 Conferences. It is false reasoning to assume that statistically significant coefficients in one sample will be significant in another same. Therefore, we cannot infer anything about College Football wins with other samples. This paper has reported the estimation of the Power Five conferences wins through a regression model. This paper presents the predictive quality only of past season data, but also a wide range of other explanatory variables. The key for the model is making an accurate prediction for the expected number of wins. The data above shows how the model was ran and what factors we took into account for the model. The model can be so far off or insanely correct which makes estimating it so interesting. The significance of each win dealt with pass yards, rushing yards, rushing defense, passing defense, and penalty yards. Each statistic had an impact on the prediction of wins. Each additional increase in pass yards, rush yards, rush defense and pass defense increase the teams chances of winning by almost 1.0 percentage points. The goal for college football teams is to win as many games as possible; this model if applied correctly can guide college football teams toward success. As shown in the model, Clemson won 9.54 regular season games while in actuality Clemson won 10.8 games. The regression doesn't take into account home field advantage or clutch tendencies for super star players. Clemson could have won that extra game due to being at home.

V. Work Cited

Watterson, John Sayle. *College football: History, spectacle, controversy*. JHU Press, 2002.

Humphreys, Brad R. "The relationship between big-time college football and state appropriations for higher education." *International Journal of Sport Finance* 1.2 (2006): 119-128.

"College Athletics - History Of Athletics In U.s. Colleges And Universities." *Sports, Intercollegiate, University, and Football - StateUniversity.com*. N.p., n.d. Web. 18 Apr. 2017.

Ridpath, B. David. "The College Football Playoff And Other NCAA Revenues Are An Exposé Of Selfish Interest." *Forbes*. Forbes Magazine, 17 Jan. 2017. Web. 18 Apr. 2017.