Predicting the Success of College Football Teams Using Multiple Regression

Brandon Kraus, Honors Student with Professor James Leininger
MidAmerica Nazarene University, Olathe, KS

Abstract

The idea behind this project was to develop an accurate and enduring formula for predicting the success of college football teams throughout the course of a college football season. Contrary to it's NFL counterpart, college football is extremely difficult to predict, particularly at the beginning of a season. For instance, Auburn University's football team had a 3-9 record in the 2012 season, including an 0-8 record in their conference games. Incredibly, Auburn put together a 12-2 record in the 2013 season, nearly winning the national championship. This variance poses problems to prognosticators across the country. Given my interest in college football, I wanted to see if I could find a statistical model that could predict the success of college football teams more effectively than my peers.

Research Questions

- What factors play a significant role in a team's success?
- Should postseason play be included?
- Is there a correlation between common factors that could improve my results?
- Is my model accurate for previous seasons or just the 2013 season?

Original Factors

There were16 factors in consideration for the formula to predict a team's win total:

- Points scored per game
- Points allowed per game
- Yards gained per game
- Yards allowed per game
- Passing yards gained per game
- Rushing yards gained per game
- Passing yards allowed per game
- Rushing yards allowed per game
- Percentage of third downs converted
- Percentage of thirds downs converted by opponent
- Turnovers caused – Turnovers committed
- Penalty yards per game
- Percentage of time scoring in red zone
- Yards gained per play
- Yards allowed per play
- How good the opponents were

Research Method: Regression Analysis

- Forward Regression: Using a software program called SPSS, variables were combined one at a time based on how strongly they were correlated to each other. This type of regression produced four factors which were necessary for determining a team's win total during the course of a season: points scored per game, points allowed per game, penalty yards per game, and turnovers caused – turnovers committed.

- Backward Regression: Beginning with the full list of variables, one variable was removed at a time based on it's P-Value until the remaining variables had a P-Value < 5%. The two necessary factors for backward regression were points scored per game and points allowed per game.

- Stepwise Regression: Taking advantage of a combination of backward and forward regression, one variable can be added and another removed simultaneously if the improvement is not significant. The three necessary factors for stepwise regression were points scored per game, points allowed per game, and turnovers caused – turnovers committed.

Each type of regression created a unique model. Therefore, it was necessary to find a way to compare the models and determine which one was the best fit. An F-test was conducted (with an alpha of .05) to decide whether the "full" model created a significant difference in the accuracy of the project, or whether the "reduced" model was adequate. When comparing the model with four factors and the model with three factors, it was discovered that there was not a substantial difference between the two. Therefore, the "reduced" model (with three factors) was chosen as the better option.

A similar process was enacted to compare the model with three factors and the model with two factors, where it was clear that there was no noticeable difference between the effectiveness of the two. Therefore, the model with two factors (points scored per game and points allowed per game) was chosen as the model that was the best fit.

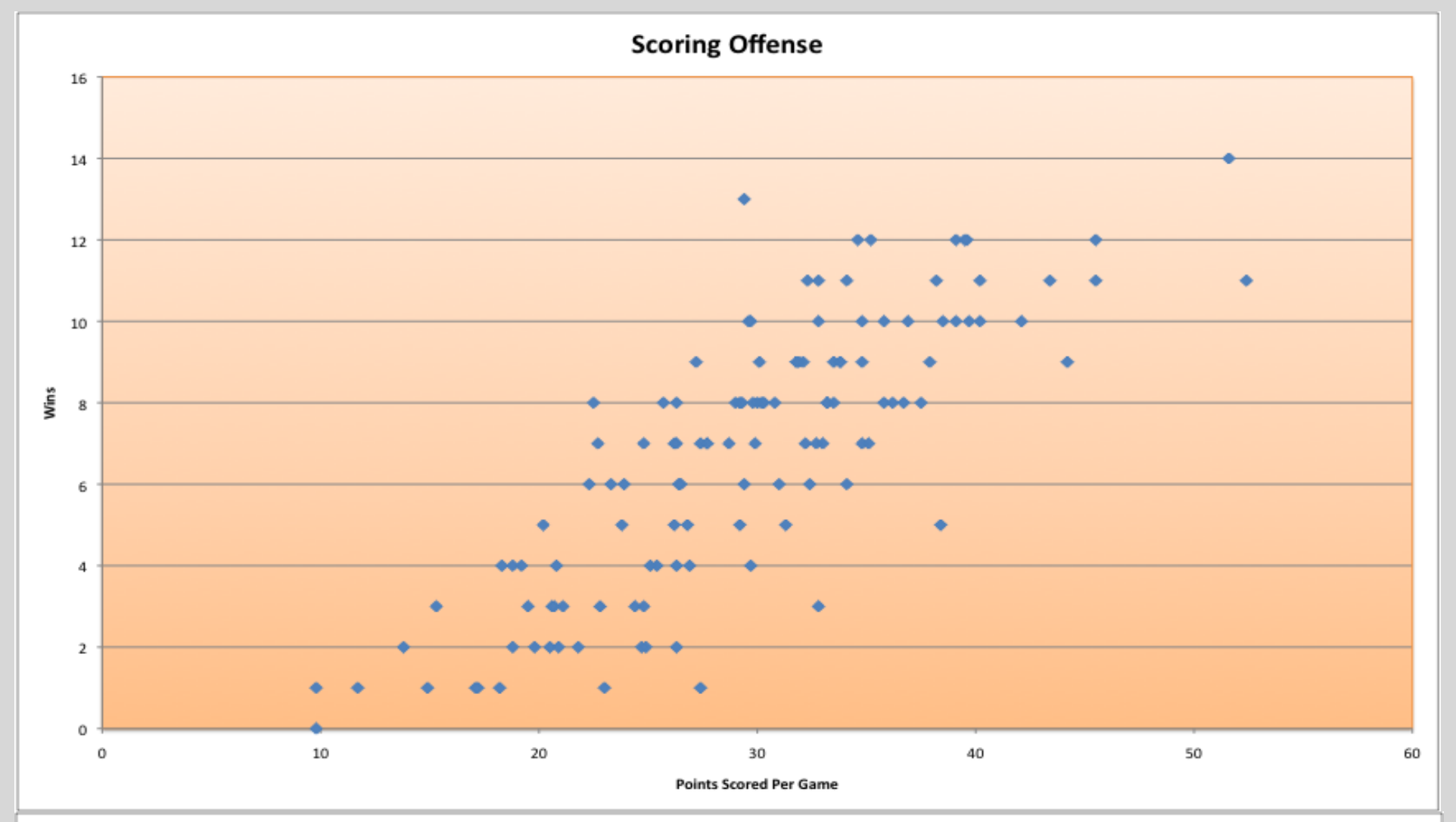
Model Formula

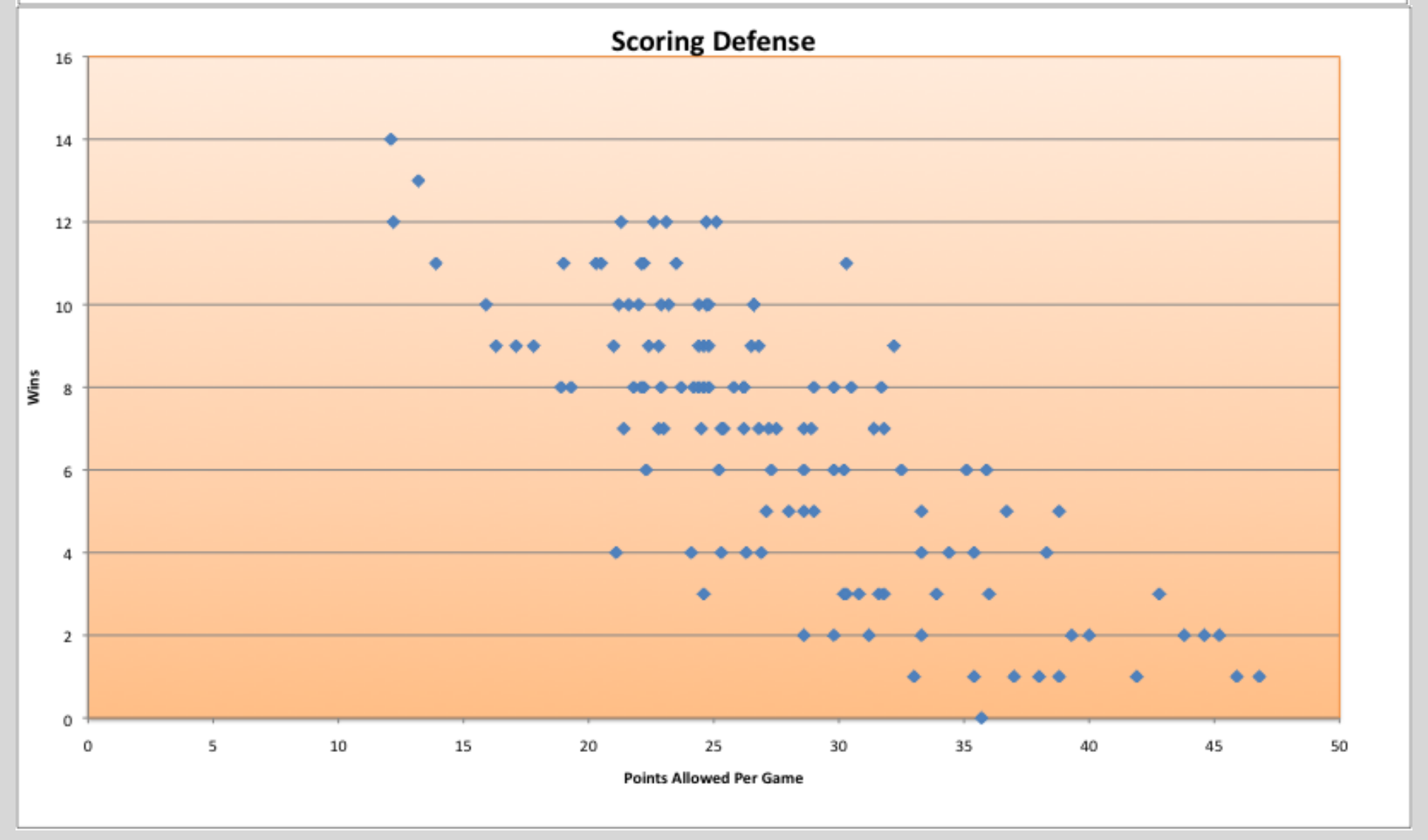
$$\text{Wins} = 6.166029157 + .240664037x - .23474558y$$

x: Scoring Offense y: Scoring Defense

Research Results: Scatter Plots

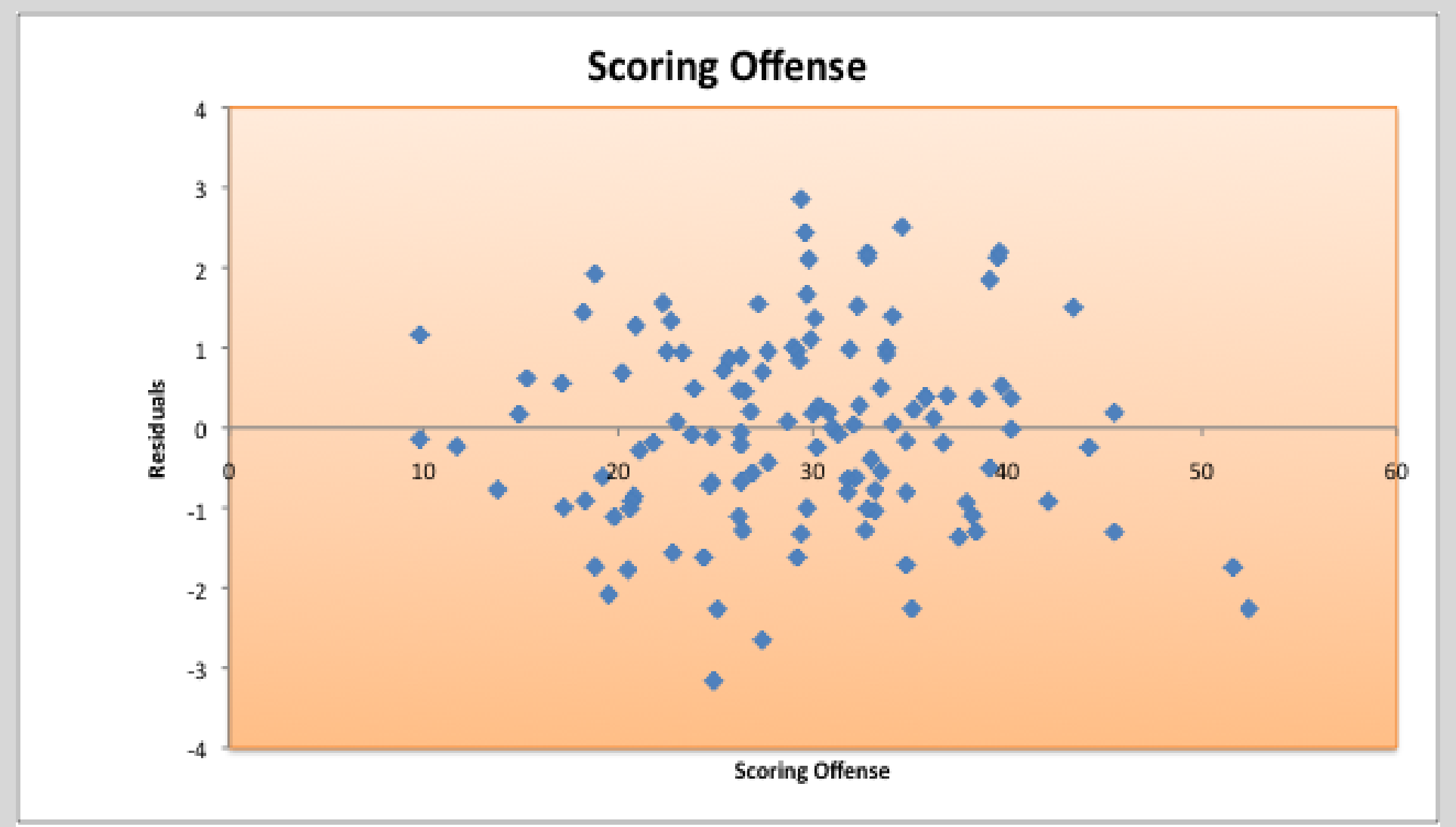
Scatter plots: Show the linear trend between the dependent variable (wins) and the two independent variables (scoring offense and scoring defense). As expected, teams that score many points and give up few points tend to win more games.

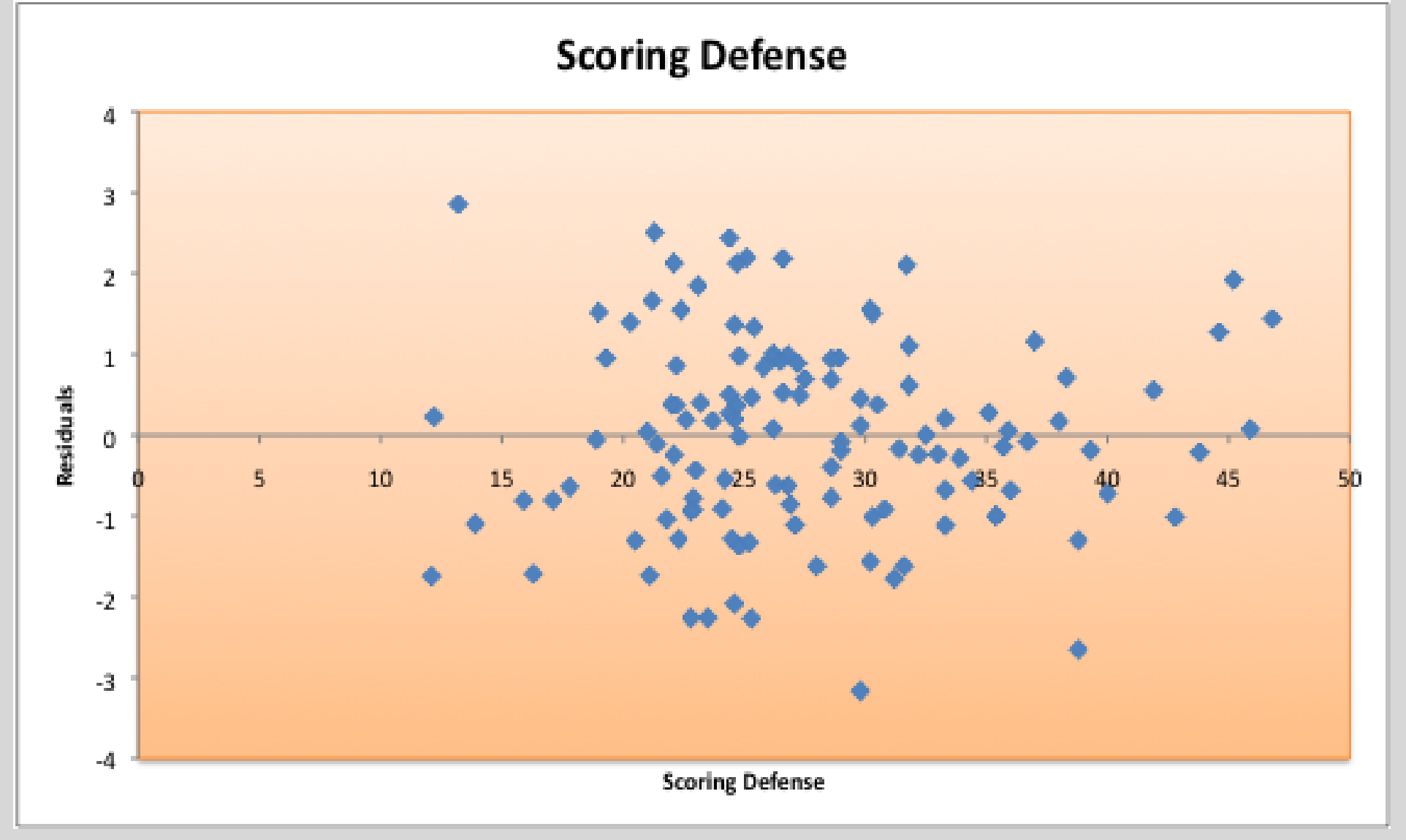




Research Results: Residuals

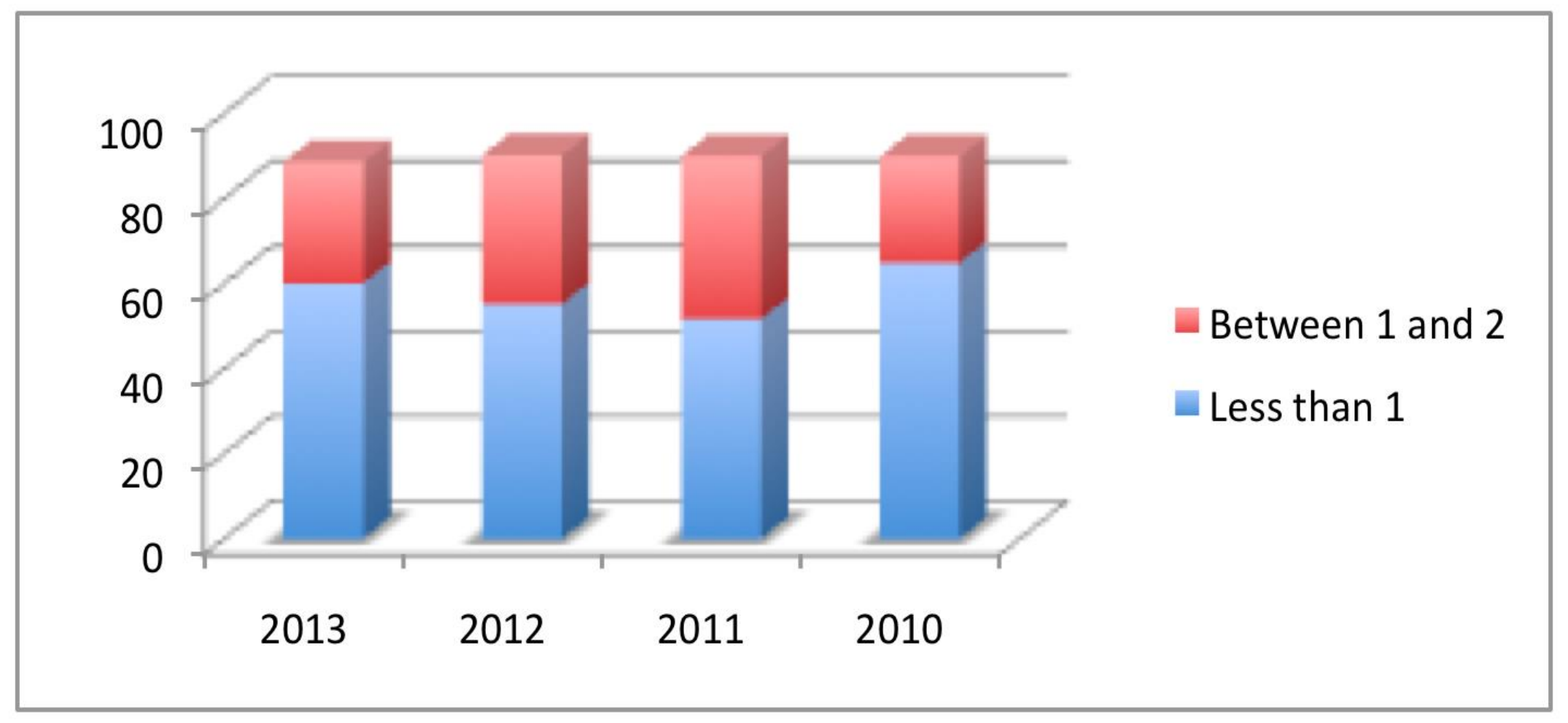
Residuals: Show the difference between the number of observed wins and the number of predicted wins based on the model. Look to see how the data fits and whether there is a pattern created. Notice that the data is clumped evenly around the line, and does not need to be fit with (i.e. does not take the form of) a higher degree function, such as a parabola or hyperbola.





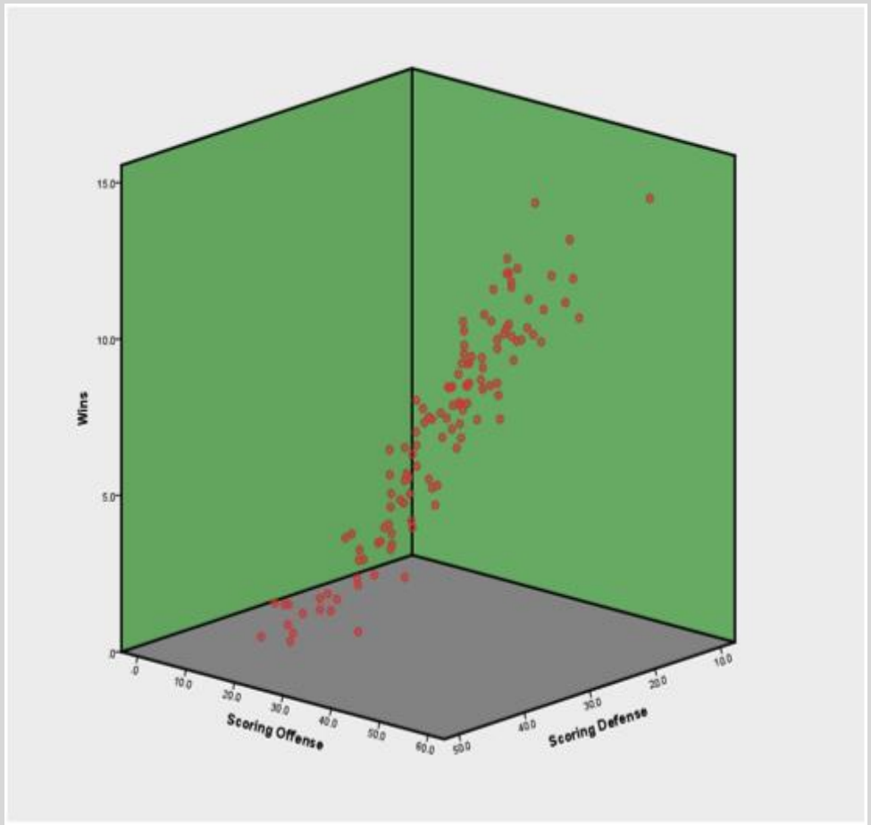
Research Results: Histogram

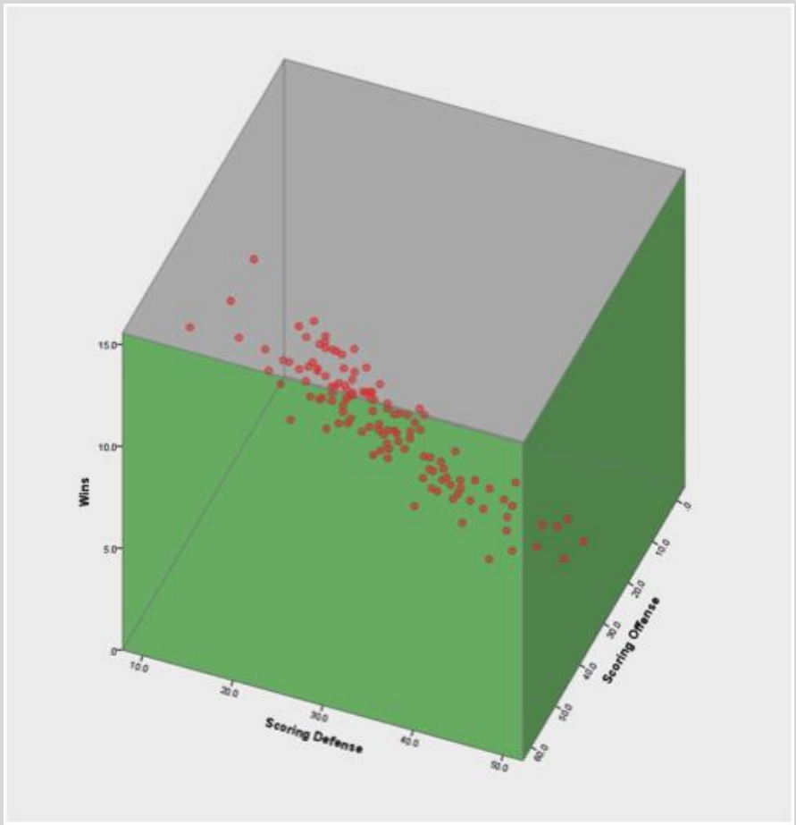
Histogram: Displays the percentage of teams whose records were predicted within one and two games of their actual record using the developed formula (around 90% each year).



Research Results: Three-dimensional Line

3D Curve: Gives a three-dimensional perspective of the correlation between the data points. It is easy to see from this view of the data that there are no outliers (points that are a significant distance from the others, or those that lie outside of the overall pattern of the data). The line can be fit with the equation listed below.





Conclusions

After running dozens of simulations, I came to the (slightly disappointing) conclusion that there are only two statistics necessary to accurately predicting season-long results for college football teams: scoring offense and scoring defense. Using these two statistics, I was able to predict the win totals of approximately 90% of the teams in NCAA division 1 within two games of their actual win totals for the 2010-2013 seasons. Given the difficulty that my predecessors have had with this task, as well as the countless factors in the life of a college football team that are unable to be quantified, I was happy with the accuracy of my results. Future study could expand on my findings by developing a formula useful in predicting results on a game-by-game basis, rather than the results of an entire season. In addition, this formula could be tested to determine whether it is an accurate representation of other levels of football, such as the National Football League.

Bibliography

(2011). NFL & NCAA Football Prediction using Artificial Neural Networks. *2011 Midstates Conference on Undergraduate Research in Computer Science and Mathematics, 1. Retrieved February 20, 2014, from http://personal.denison.edu/~lalla/MCURCSM2011/4.pdf*

Reid, M. (2003). Least Squares Model for Predicting College Football Scores. Colosseum, 1. Retrieved February 7, 2014, from http://www.geocities.ws/Colosseum/Gym/3270/bcs.pdf

College Football Stats. *TeamRankings*. Retrieved January 29, 2014, from http://www.teamrankings.com/ncf/stats/