BAS 320 - Assignment 3 - Associations Part 2

Jack Guptill

## Introduction and findings

I have been a lifelong fan of collegiate football. I even played as a defensive end in high school. One of my old coaches used to always say that defense wins games. With this report I will be seeing if that statement is somewhat true or maybe why. For this analysis we will be looking at the data from a defensive perspective and seeing if the variables chosen have any correlation to wins or losses.

The first association I want to analyze is the relationship between the number of yards allowed per game and wins for college football teams. The amount of yards allowed is the total amount of yards the defense gives up when the offense runs or passes the ball. I am assuming that the relationship is significant, but in this report, I will be able to tell you what the degree of the relationship is and what percentage of variation can explain the outcome (win/loss).

After some analysis, this association turns out to be statistically significant with a p-value of less than .05, implying that there is a relationship between the total amount of yards allowed and wins that cannot be explained by random chance. In other words there is significance to the relationship between the two variables.

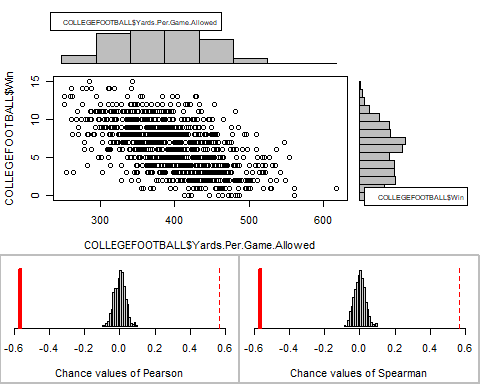
My second assication I want to analyze is the relationship between the percentage of points allowed from the Red-Zone (20 yards and in) and total losses. The red zone is often seen as either a strong suite or a pitfall of certain teams due to the defensive schemes utilized. I am curious if we can explain varitation in losses by the percentage of points allowed within the red zone.

This association also turns out to be statistically significant with a p-value of less than .05, implying that there is a relationship between the percentage of points you allow in the red zone and the number of losses a team will have that cannot be explained by random chance.

## Part 1: Analysis of Yards Allowed Per Game and Total Wins

The data I’m using comes from the BAS 320 Fall 2023 Dataset and contains a total of 970 rows with 145 columns.

A graph with green dots and red line

Description automatically generated

The relationship between wins and the number of yards allowed per game is negative and linear. The stream of points has roughly equal vertical spread in the amount of wins for every given yard allowed which tells us that the data is homoscedastic.

The relationship is moderate based on the pearson r value being -.56 and the general scatterplot is an elipse shape following a general negative trend line. Because of homoscedasticity and the data following a monotonic linear trend, the pearson correlation must be used to study the association. We can determine from the r value stated earlier that the r^2 value equals .317. In other words, 31.7% of the variation in wins can be explained by the number of total yards allowed per game. By knowing the number of yards allowed by a team, you will have 31.7% of the information needed to determine how many wins that team will have.

Based on the estimated p-value of 0, the relationship is statistically significant. The practical implication of this association is that high winning teams tend to have a lower number of yards allowed. While there is not a cause-and-effect relationship between the variables, some of the variation in wins is explained by the total yards allowed as mentioned earlier.

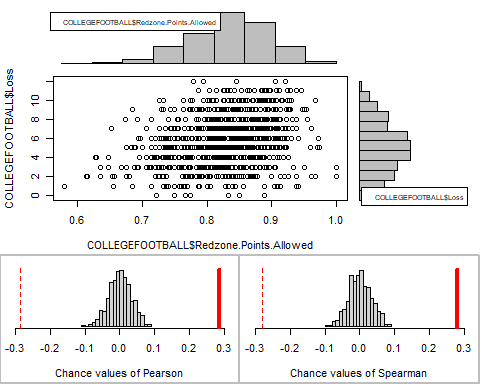
## Part 2: Analysis of Percentage of Total Points in the Red Zone and Total Losses

The relationship between losses and the percentage of redzone points allowed is positive and linear with a monotonic direction. The stream of points has roughly an equal vertical spread in the number of losses for every percent increase in red zone points allowed, which tells us that the data is homoscedastic.

The relationship is weak based on the Pearson’s r correlation which is .29. Additionally, the scatterplot shows an ellipse shape for the data that is following a rough positive trend line. Because of homoscedasticity and the data following a monotonic linear trend, the Pearson’s r correlation must be used to study the association. We can take the r value of .29 and calculate the r^2 value which equals .084 or 8.4%. We now know that 8.4% of the variation in losses can be explained by the percentage of red zone points allowed by a team.

Based on the estimated p-value of 0, the relationship is statistically significant. The practical implication of this association is there tends to be a lower percentage of overall points from the red-zone with winning teams. While there is not a cause-and-effect relationship between the variables, some of the variation in losses is explained by points scored from the red-zone.

A graph of red and blue points

Description automatically generated

## Conclusions/Next Steps

Based on the findings and analysis done, my coach was somewhat right based on what we looked at. We now know that there is a relationship, no matter if it is small or somewhat large, between the variables analyzed. It also seems that defense plays a big role in not just wins, but also in losses.

Things you might want to prioritize for the future, from a coaching/scheme perspective would be to minimize total yardage allowed within games. This doesn’t explain everything when comes to wins, but there is a relationship there. You also need to focus on stopping the run or pass within the redzone to hopefully increase your likelihood of not losing.

Going forward it would be beneficial to look for more relations to both wins and losses to try and determine which variables explain the most in terms of variation. A correlation matrix would be a good next step.