Can You Predict the Number of Wins in a season in the NFL based on a previous season?

Avinash S. Perera

St. Petersburg College

STA 2041: Data Analysis & Statistical Modeling

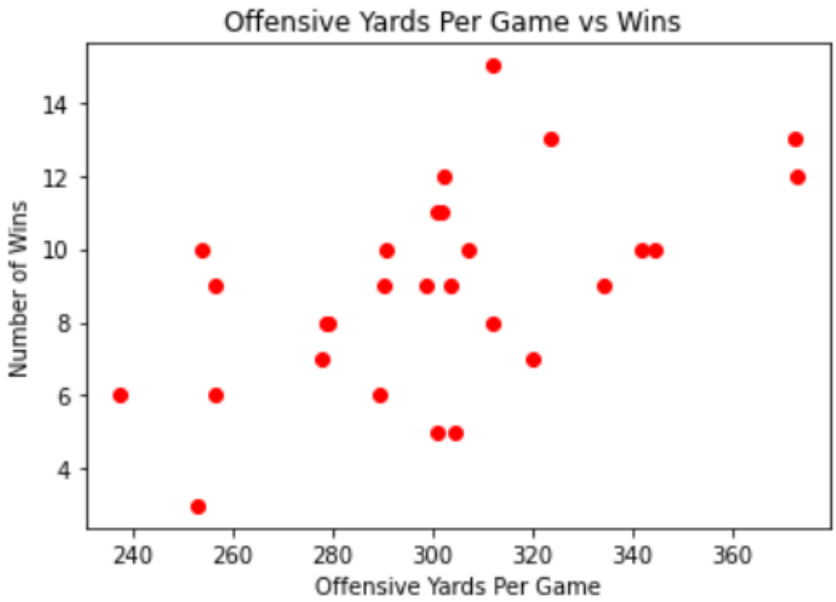
E. Gretchen Gascon

April 29, 2023

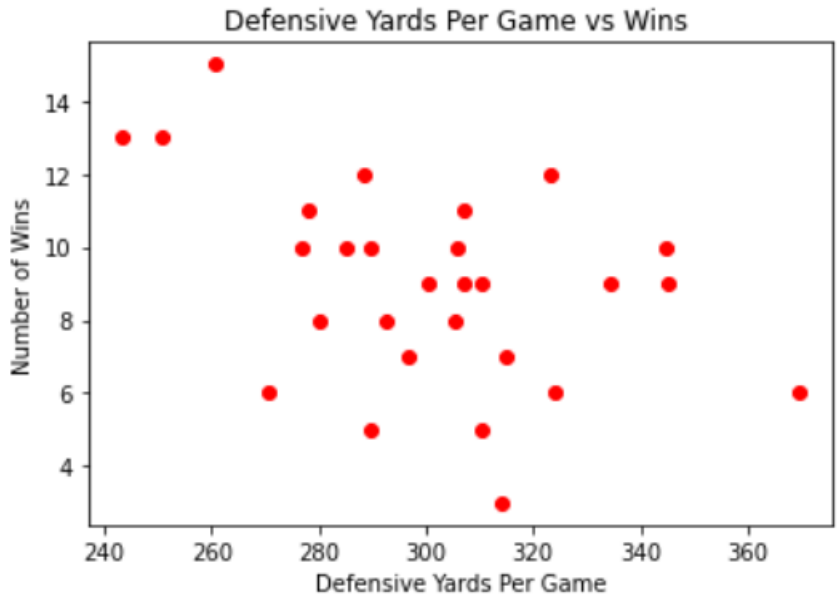
Introduction

The data used is a database of NFL team statistics from a previous season. We will begin by assuming that all the requirements of a Multiple Linear Regression model are being met. We will attempt to create a Multiple Linear Regression model that uses team statistics like offensive yards per game or defensive points per game to most effectively predict the number of games won by an NFL team. Scatterplots will be used to visualize the correlation between all the independent variables. Then a Simple Linear Regression model will be used to predict wins using the number of Defensive Yards given up per game. Then a Multiple Linear Regression model will be used to predict wins using Defensive Yards given up per game and Offensive Yards taken per game. Then a model will be created that incorporates all four predictor variables. Finally, all permutations of the predictor variables will be tested in models to choose the best one for predicting wins. In choosing the best model, the predictor variables' correlations with each other will be considered, as well as t-tests for the coefficients of the regression equation, and finally the adjusted coefficients of determination for each possible model will be considered.

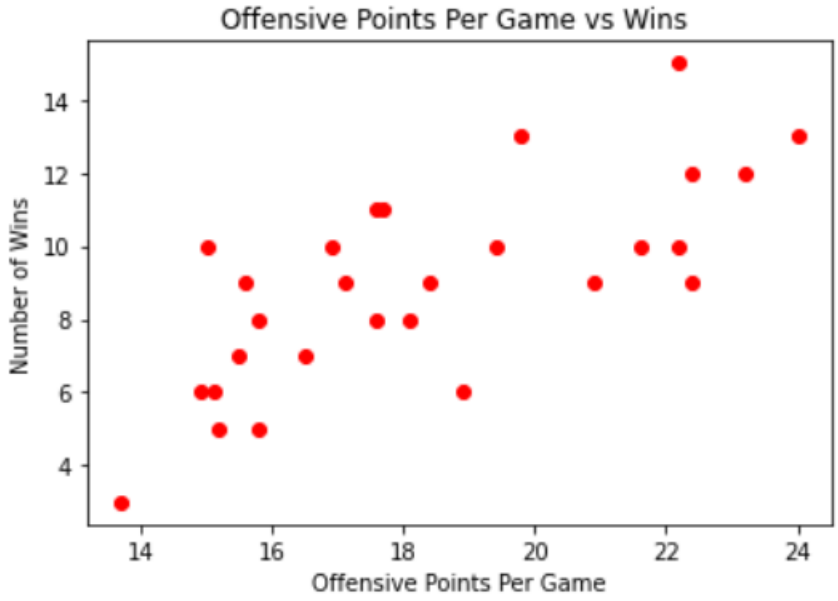
Data Visualization



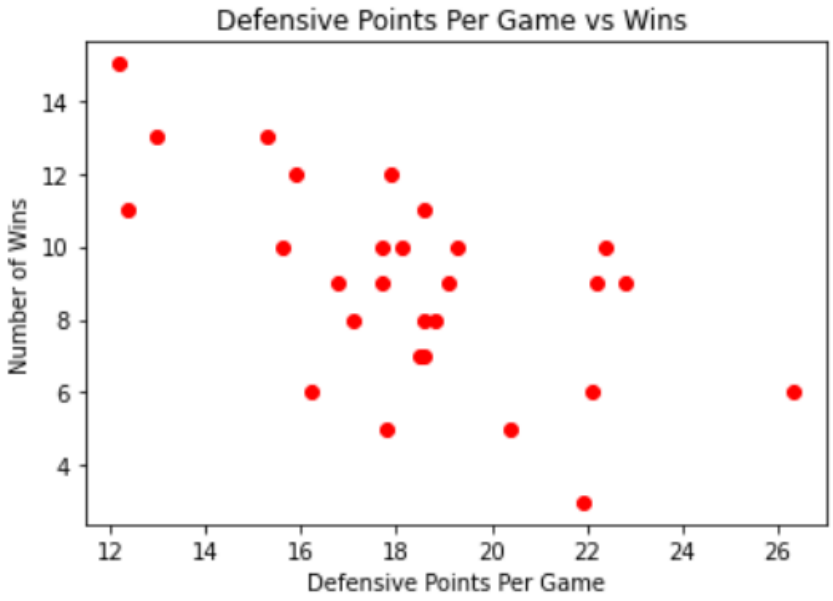

```
Correlation between Offensive Yards scored and the Total Number of Wins
Pearson Correlation Coefficient = 0.5482
P-value = 0.0025
```



Correlation between Defensive Yards scored and the Total Number of Wins
Pearson Correlation Coefficient = -0.4373
P-value = 0.02



Correlation between Offensive Points scored and the Total Number of Wins
Pearson Correlation Coefficient = 0.7235
P-value = 0.0




```
Correlation between Defensive Points scored and the Total Number of Wins
Pearson Correlation Coefficient = -0.6054
P-value = 0.0006
```

These scatterplots give us an idea of how a predictor variable correlates to the response variables. This helps us understand how useful it may be in the regression model. In general, we see that yards per game for both sides have a weaker correlation to wins than points per game. In general, we see that defensive statistics are negatively correlated to wins and offensive statistics are positively correlated to wins. The strongest correlation to wins, regardless of direction, are offensive points per game. This intuitively makes sense since the most offensive points in a game causes a win, so wins will closely follow the number of offensive points per game. Since all p-values are lower than the alpha of 0.05, all the correlation coefficients are statistically significant.

Simple Linear Regression: Predicting Wins Using Defensive Yards Per Game

A Simple Linear Regression model works by using a linear equation of regression. Point estimates & confidence intervals are calculated for the y-intercept & the slope of the equation. $\hat{Y}$ is the predicted value of the response variable, and the coefficient of the x-term or slope is the value of the predictor variable. The slope represents the effect on the response variable from changes in the predictor variable.

Regression Equation: $\hat{y} = 21.62 + -0.04X_{\text{Def. Yards}}$

F-test:

H_0: $\beta_1 = 0$ (The independent variable has no statistically significant effect on the response variable)

H_a: $\beta_1 \neq 0$ (The independent variable has some statistically significant effect on the response variable)

$\alpha = 0.05$

F: 6.148, P-value: 0.02

Since the p-value is less than the alpha of 0.05, we must reject the Null Hypothesis. There is evidence to support that the predictor variable of Defensive Yards given per game has some effect on the number of wins.

Given a Defensive Yards per game of 290, the predicted number of wins is approximately 9.
21.6243 + (-0.0421 x 290) = 9.4153

With a coefficient of determination of 0.191, about 19% of the variation in number of wins is explained by variation in defensive yards per game. It gives us some idea of how well the model fits the data. Since this model only uses one predictor variable, we will use the regular R^2 instead of the adjusted one.

Multiple Linear Regression: Predicting Wins Using Defensive Yards Per Game & Offensive Yards Per Game

A Multiple Linear Regression model works similarly to a Simple Linear Regression model, except more than one predictor variable is introduced. The model still works based off of a linear equation, it just has more x-terms. The slope of each x-term shows how the response variable is affected by changes in each respective predictor variable.

Regression Equation: $\hat{y} = 7.11 + -0.03X_{\text{Def. Yards}} + -0.04X_{\text{Off. Yards}}$

Adjusted R^2: 0.372 (Roughly 37% of the variation in number of wins is explained by the variation in the defensive & offensive yards per game)

F-test:

H_0: $\beta_1 = \beta_2 = 0$ (The independent variables have no statistically significant effect on the response variable)

H_a: $\beta_1 \neq 0$ or $\beta_2 \neq 0$ (At least one independent variable has some statistically significant effect on the response variable)

$\alpha = 0.05$

F: 8.985, P-value: 0.00

Since the p-value is less than the alpha of 0.05, we must reject the Null Hypothesis. There is evidence to support that at least one of the predictor variables has some effect on the number of wins.

t-tests:

Defensive Yards:

t: -2.25, p-value: 0.03, reject null since p-value is less than alpha of 0.05, Defensive Yards has some effect on wins.

Offensive Yards:

t: 3.12, p-value: 0.00, reject null since p-value is less than alpha of 0.05, Offensive Yards has some effect on wins.

Given a Defensive Yards per game of 290 & Offensive Yards per game of 307, the predicted number of wins is approximately 10.

7.1130 + (-0.0336 x 290) + (0.0398 x 307) = 9.5876

Multiple Linear Regression: Predicting Wins Using Defensive Yards Per Game, Offensive Yards Per Game, Offensive Points Per Game, Defensive Points Per Game

Regression Equation: $\hat{y} = 5.67 + 0.00X_{\text{Def. Yards}} + 0.00X_{\text{Off. Yards}} + 0.60X_{\text{Off. Points}} + -0.44X_{\text{Def. Points}}$

Adjusted R^2: 0.703 (Roughly 70% of the variation in number of wins is explained by the variation in all four predictor variables, this has changed from the previous model)

F-test:

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ (The independent variables have no statistically significant effect on the response variable)

H_a: At least one $\beta_i \neq 0$ for i = 1, 2, 3, 4 (At least one independent variable has some statistically significant effect on the response variable)

$\alpha = 0.05$

F: 16.98, P-value: 0.00

Since the p-value is less than the alpha of 0.05, we must reject the Null Hypothesis. There is evidence to support that at least one of the predictor variables has some effect on the number of wins.

t-tests:

Defensive Yards:

t: 0.24, p-value: 0.81, fail to reject null since p-value is greater than alpha of 0.05, Defensive Yards have no effect on wins.

Offensive Yards:

t: -0.23, p-value: 0.82, fail to reject null since p-value is greater than alpha of 0.05, Offensive Yards have no effect on wins.

Defensive Points:

t: -2.68, p-value: 0.01, reject null since p-value is less than alpha of 0.05, Defensive Points has some effect on wins.

Offensive Points:

t: 3.79, p-value: 0.00, reject null since p-value is less than alpha of 0.05, Offensive Points has some effect on wins.

Based on this, only Defensive & Offensive Points should be used in the model.

Given a Defensive Yards per game of 290, Offensive Yards per game of 307, Offensive Points per game of 17, and Defensive Points per game of 20, the predicted number of wins is approximately 7.

5.6673 + (0.0045 x 290) + (-0.0033 x 307) + (0.6043 x 17) + (-0.4439 x 20) = 7.3543

Conclusion

This analysis has shown that the ideal model for predicting player wins based on this data has Defensive Points & Offensive Points as predictor variables. Defensive & Offensive Yards have higher Multicollinearity. Yards predictor variables also have weaker correlations to wins than the Points predictor variables. The scatterplots and the correlation table of the dataframe show this. During the individual t-testing of the model that includes all four variables, we found that only Defensive Points & Offensive Points have evidence supporting that they have some effect on the response variable. Finally, when the adjusted R^2 of each possible combination of predictors was calculated, the model that only included Defensive & Offensive Points had the highest R^2 value (0.725). This means that this particular model has the most variation in wins explained by the variation in predictor variables. Therefore, the most effective predictor variables for modeling wins for an NFL team are Defensive Points and Offensive Points. It is possible to somewhat predict the next NFL season's statistics based on a previous season's data Some other data being collected by the NFL that could help predict future wins per team are statistics such as number of Super Bowl wins per team, average player salary, and number of player injuries for the season in question.