Dylan Johnson
03/10/2023
DA 401, Spring 2023
Professor Matt Lavin
Early Results Assignment

"Separating Good from Great: Analyzing the Impact of Separation on an Individual NFL Receiver Success"

My research question is "Does an NFL Receiver's Separation lead to success in the NFL?”. I want to see whether or not an NFL receiver's separation rate is directly linked to their success in the NFL. The separation rate is a measure of how much distance a receiver creates from their defender during a play. I hypothesized that receivers with higher separation rates will have more success in the NFL, as they are more likely to get open and catch passes from their quarterback. To test this hypothesis, I have collected data on receiver separation rates and their performance metrics, such as yards per reception and touchdowns, and analyze the relationships between the variables.

I collected my data from the Next Gen Stats website, which provides comprehensive information on NFL receiver performance, including their separation rate, yards gained, receptions, touchdowns, and other related metrics. I will be focusing on the last five seasons worth of data. To analyze the relationship between separation rates and success in the NFL, I have decided to use regression analysis. Regression analysis will help me build a predictive model that identifies the relationship between separation rates and various performance metrics, while controlling for other factors that may influence receiver success, such as the quality of the quarterback and opposing defense. My choice of method is consistent with established research practices in sports analytics. For instance, previous research has shown that regression analysis is commonly used to model the relationship between performance metrics in sports (Grier & Altman, 2018), while hypothesis testing is often used to examine the significance of differences in performance between different groups of athletes (Bock, 2013).

For my initial model I started with the regression analysis between separation and all of the other receiving statistics in my dataset. So, I started by testing the regression between 'SEP' and all other statistics. What I found was that the relationships between variables were actually very surprising. In this case, the F-statistic of my entire model is 84.36 with a p-value of 1.02e-108, indicating that the model is statistically significant. There was also a lot of variation in the results and in Figure 1 you can see the basic relationship plots of each variable. Based on the results from the first test, the relationship between 'SEP' and the following variables are displayed in Figure 2.

Figure 1: Regression Relationships of 'SEP' to Dependent Variables

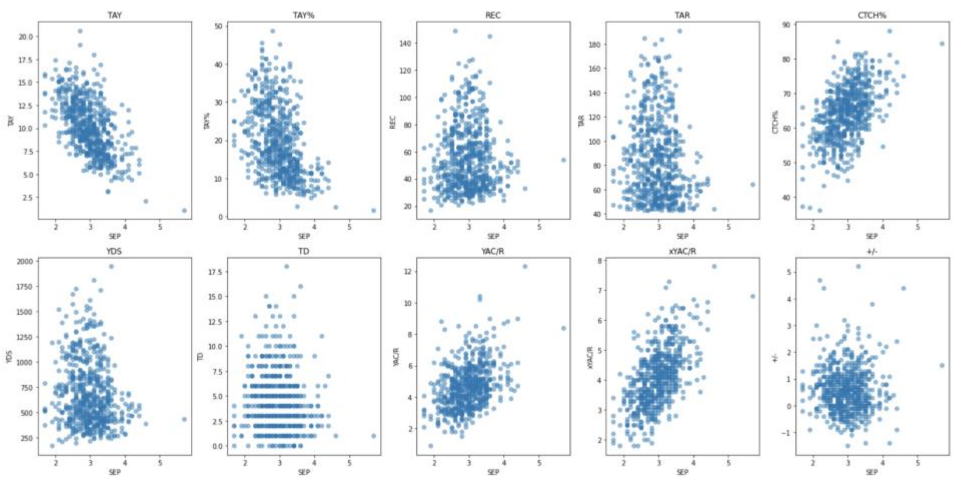

Figure 1: *In this plot chart, TAY, TAY%, REC, TD, YAC/R, xYAC/R, and +/- are not statistically significant predictors of SEP, as their p-values are greater than 0.05. However, TAY%, REC, TD, YAC/R, xYAC/R, and +/- may still have some predictive power in combination with other variables in the model.*

Figure 2: Regression Relationship Analysis of 'SEP' to Dependent Variables

Dependent Variable	Relationship
TAY	Negative
TAY%	Negative

Dependent Variable	Relationship
TAY	Negative
REC	Negative
TAR	Positive
CTCH%	Positive
YDS	Negative
TD	Positive
YAC/R	Positive
xYAC/R	Negative
+/-	Negative

Figure 2: *In this figure it is evident that 'TAR', 'CTCH%', 'TD', and 'YAC/R' all have positive relationships to 'SEP' whereas 'TAY', 'TAY%', 'REC', 'YDS', 'xYAC/R', '+/-' all have negative relationships. Therefore, we can see that my data is showing an increase in a receivers Separation (SEP), will also increase their Targets (TAR), their Catch Percentage (CTCH%), their number of Touchdowns (TD), and the (YAC/R).*

All of the results will help me further my analysis, but I wanted to point out a few of the results that will directly impact my next move. I wanted to individually check the regression for each of these dependent variables individually, in order to see how much each variable is affected by 'SEP'. I was able to do this by focusing variables such as that R-squared values; which indicate the proportion of the variance in 'SEP', the P-values; which indicate the statistical significance of my variables, the F-statistic; which tests whether the independent variables as a group are significant predictors, the Standard Error; which provide information about the precision of the estimates, and the coefficients; which gives me the relationships between the variables.

First, I looked at the relationship between separation (SEP) and Total Air Yards (TAY). For this regression 'SEP' is the independent variable and 'TAY' is the dependent variable. The regression equation goes like this, TAY = 20.1488 - 3.3822(SEP). The R-squared value of 0.339 indicates that 33.9% of the variation in the Total Air Yards can be explained by the receiver's separation. The p-value of less than 0.05 suggests that it is unlikely that the observed relationship between SEP and the dependent variable 'TAY', is coincidental. The F-statistic of 322.5 with a corresponding p-value of 1.58e-58 indicates that the overall model is significant. The F-statistic tests the overall significance of the regression model and the very low p-value (1.58e-58) indicates that the probability of observing such a large F-statistic by chance is extremely small, providing strong evidence to reject the null hypothesis that the model is not significant. The following scatter plot shows a clear negative relationship between 'SEP' and 'TAY', indicating that as 'SEP' increases, 'TAY' decreases.

Figure 3

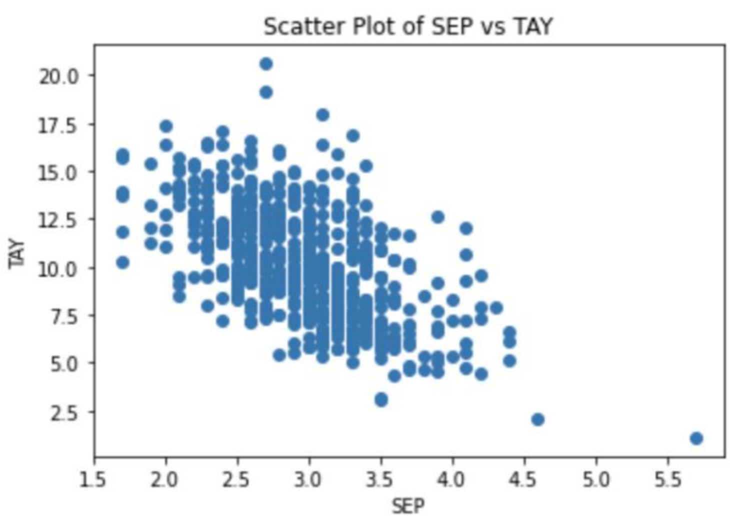

Figure 3: The coefficient for the independent variable 'SEP' is -3.3822, which means that for every unit increase in 'SEP', 'TAY' decreases by 3.3822 yards. Overall, the early results from my model suggest that 'SEP' is a significant predictor of 'TAY' in the given dataset, and that the 'TAY' goes down for receivers who get more separation. Therefore, receivers tend to get the ball faster when they get more separation.

Next, I want to dive into the Catch Percentage (CTCH%) variable. The regression analysis shows that there is a statistically significant positive relationship between CTCH% and SEP (separation). The coefficient of determination (R-squared) is 0.26, which means that 26% of the variation in CTCH% can be explained by the variation in SEP. The p-value for the slope coefficient is less than 0.05, indicating that the relationship between CTCH% and SEP is significant. The intercept term of 41.14 indicates the expected value of CTCH% when SEP is equal to 0. The regression line has a positive slope of 7.87, which means that as SEP increases, CTCH% also tends to increase. The omnibus test and Jarque-Bera test are not statistically significant, indicating that the residuals are normally distributed. Overall, these results suggest that SEP is an important predictor of CTCH%, and players with higher separation tend to have a higher catch percentage.

The last variable I wanted to mention was the Targets (TAR), because the results were not what I was expecting. The regression analysis shows that the independent variable "SEP" has a statistically significant negative effect on the dependent variable "TAR" (t=-3.135, p=0.002). The model has a low R-squared value of 0.015, indicating that only a small portion of the variance in "TAR" can be explained by "SEP". The scatter plot generated from the data shows a weak negative correlation between "SEP" and "TAR", with the data points spread out and not fitting a clear linear pattern. This further supports the low R-squared value and suggests that other factors may be more influential in determining the number of targets a player receives. The analysis also reports the coefficient of determination, indicating that only 1.5% of the variation in "TAR" can be explained by "SEP". Therefore, while there is a statistically significant relationship between "SEP" and "TAR", it is a weak relationship, and other factors may play a more significant role in determining the number of targets a player receives. Further analysis with additional variables may help to improve my model's predictive power and identify other factors that contribute to a player's number of targets.

In conclusion, my early results were very helpful in the sense that I now know each dependent variable’s contribution to my analysis. My model provides evidence that receivers have a greater chance of catching the football when they have more separation. They also tend to lose air yards as they increase their separation which was more relevant than I had originally predicted. Both of these variables are significant enough that I will carry them with me to my next task in my analysis. There are many other variables and results that I will use in my future work. Although NFL receivers with more separation may get a few extra balls thrown their way, the data provided does not give us the notion that more separation leads to more targets. In the near future, I also intend on using a hypothesis test such as a t-test or ANOVA, to assess whether there is a statistically significant difference in the performance of receivers with high and low separation rates. All in all, my early results from my regression analysis proved to be not only surprising, but insightful, and at times, extremely significant.

References

Bock, T. M. (2013). Applied statistics in sport research. Journal of Sport Psychology in Action, 4(3), 135-146.

Grier, S. A., & Altman, D. G. (2018). Statistics applied to the clinical and health sciences (5th ed.). Elsevier.

Link to GitHub Repo: git@github.com:DylanJohnson02/da-401-johnson.git