*Sydney Scaccia*

DSC 501: Introduction to Data Science

**Final Project Statistical Findings**

**Project Recap**

Throughout the past seven weeks the research question under investigation continued to be tweaked, but the independent and dependent variables have stayed the same. Both the variables and the research question are computed from data that was collected throughout the whole season as opposed to game-by-game statistics. The dependent variable is the team’s overall performance throughout the season and will be evaluated by using the team’s win percentage. The independent variables are yards allowed, points allowed, points gained, rush yards and pass yards which will combine to give the total number of yards gained throughout the season. Last week points allowed and points gained were found to be statistically significant variables. The question under investigation is: **Does defense have an impact on team winning percentage?** The null hypothesis is defense does not impact win percentages and the alternative hypothesis is defense does impact a team’s win percentage. In last week’s analysis the null hypothesis was rejected based off the statistical findings.

**Analysis / Model**

Throughout last week’s statistical analysis, a linear regression model was used along with a stepwise model to reiterate the outputs and to ensure accuracy. The statically significant variables, total points and points allowed, had potential for multicollinearity when used with the dependent variable of win percentage. To combat this, prior to implementing the stepwise function, an association analysis was added to detect the correlation between the variables. Table one is the output after running the association analysis. The independent variable, total points, is highly associated with the dependent variable, win percentage.

**Table One: Pearson Association Analysis – Win Percentage**

|  |  |  |
| --- | --- | --- |
|  | **Association Measure** | **p-value** |
| **Total Points** | **0.81955** | **0** |
| **Points Allowed** | **-0.64226** | **2.2204E-16** |
| **Rush Yards** | **0.5152** | **3.5701E-10** |
| **Pass Yards** | **0.43081** | **3.1133E-07** |
| **Yards Allowed** | **-0.38595** | **5.7604E-06** |

Below in Table two, the full correlation matrix is displayed. Table two shows that total points and points allowed are not highly correlated which was anticipated from last week. This means there is not as much multicollinearity between the two variables. On the other hand, for total points, there is a possibility for multicollinearity with the dependent variable because the more points a team scores throughout the season ultimately increases their chances of winning games leading to the team having a better win percentage. The small amount of multicollinearity present in the total points should not outshine the other variables in any further analysis.

**Table Two: Full Correlation Matrix**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Win Percentage** | **Yards Allowed** | **Pass Yards** | **Rush Yards** | **Points allowed** | **Total Points** |
| **Win Percentage** | **1** | **-0.385951** | **0.43813** | **0.515203** | **-0.642265** | **0.819554** |
| **Yards Allowed** | **-0.385951** | **1** | **0.147915** | **-0.19604** | **0.833855** | **-0.108137** |
| **Pass Yards** | **0.43813** | **0.147915** | **1** | **-0.22939** | **-0.041913** | **0.611258** |
| **Rush Yards** | **0.515203** | **-0.19604** | **-0.22939** | **1** | **-0.316092** | **0.562728** |
| **Points Allowed** | **-0.642265** | **0.833855** | **-0.041913** | **-0.316092** | **1** | **-0.351782** |
| **Total Points** | **0.819554** | **-0.108137** | **0.611258** | **0.562728** | **-0.351782** | **1** |

Upon completing the association analysis, a linear regression model was added producing the same outcome as last week. Once completing the linear regression, a stepwise function continued the analysis to look deeper into these independent variables. The alpha level being used to determine if an independent variable is statistically significant is .05. This will help determine if the null hypothesis should be accepted or rejected based upon each variable’s p-value. Table Three shows the linear regression output and Table Four shows stepwise regression output.

**Table Three: Linear Regression**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Min** | **1Q** | **Median** | **3Q** | **Max** |
| -30.682 | -6.97 | -0.291 | 7.217 | 23.911 |
| **Coefficients:** |  |  |  |  |
|  | **Estimate** | **Std. Error** | **T value** | **Pr(>|t|)** |
| **Intercept** | 32.22509 | 8.995846 | 3.5822 | .00049 |
| **Yards Allowed** | 0.003206 | 0.002687 | 1.1929 | 0.23518 |
| **Pass Yards** | 0.000365 | 0.002774 | 0.1315 | 0.8956 |
| **Rush Yards** | 0.000792 | 0.003155 | 0.2511 | 0.80217 |
| **Points Allowed** | -0.14096 | 0.023006 | -6.127 | 1.09E-08 |
| **Total Points** | 0.134158 | 0.02651 | 5.1498 | 9.94E-07 |
| Win % = 32.22509 + 0.003206 (Yards Allowed) + 0.00365 (Pass Yards) + 0.00792 (Rush Yards) + -0.014096 (Points Allowed) + 0.134158 (Total Points) | | | | |
| Multiple R Squared = 0.817, Adjusted R Squared = 0.8096 | | | | |

**Table Four: Stepwise Regression**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Min** | **1Q** | **Median** | **3Q** | **Max** |
| -30.07 | -7.34 | -0.23 | 7.32 | 23.36 |
| **Coefficients:** |  |  |  |  |
|  | **Estimate** | **Std. Error** | **T value** | **Pr(>|t|)** |
| **Intercept** | 39.584 | 6.05168 | 6.541 | 1.35E-09 |
| **Points Allowed** | -0.1173 | 0.011847 | -9.898 | < 2.2e-16 |
| **Total Points** | 0.1428 | 0.008605 | 16.6 | < 2.2e-16 |
| Win % = 39.584 + -0.1173 (Points Allowed) + 0.1428 (Total Points) | | | | |
| Residual Standard Error = 9.6166 on 127 degrees of freedom | | | | |
| F-statistic = 279.1 on 2 and 127 degrees of freedom (DF), p-value < 2.2e-16 | | | | |

The linear regression model, Table Three, output a multiple R squared value and an adjusted R squared value. These values do not have a significant difference between them which could be caused by the multicollinearity is baked into the original independent variables that cannot be individually detected. As previously discussed, this could be present in the “total points” independent variable. Total points and points allowed were found to be significant because the p-values for these variables are below the .05 threshold and must be investigated further. Each variable also has a large enough coefficient meaning it will ultimately impact the model.

Table Four, the stepwise model, only total points and points allowed have been included in the analysis. The multiple R squared and the adjusted R squared values do not have a large difference between them. Each variables p-value is below the .05 threshold making them statistically significant. This can now aid in the rejection or the acceptance of the null hypothesis which ultimately answers the initial research question.

**Discussion**

Once completing the statistical analysis of the independent variables, the initial research question can be answered. The null hypothesis, defense does not impact win percentages, can now be rejected because the p-values that were found are statistically significant. The alternative hypothesis, defense does impact a team’s overall win percentage, can be accepted. This could be considered a Type I error because the null hypothesis has the potential of being wrongfully rejected (Wheelan, 2014). Based off outside literature this has a low potential of being a Type I error.

The article, *“Defense Win Championships” Proves True in Hoops*, written by Tracy Pierson (2018), supports the statistical findings that defense does impact a team’s overall win percentage. Pierson focuses her discussion on NCAA basketball by examining how having a solid defensive effort will alter the game’s outcome. Pierson states the key to winning championships and games in general is to have a team that plays disciplined defense; the most successful teams come from teams that have their top players play a well-rounded game (2018).

*Does Defense Win Championships?* displays both sides of this controversial argument (The Stats Zone, 2016). The Stats Zone found that playing with a solid defense on the court increases the team’s win percentage by 2.1 percent as opposed to having a strong offense alone (The Stats Zone, 2016). However, this article also shows there is room for error, or outliers, such as successful defenses losing games they have been predicted to win and vice versa (The Stats Zone, 2016).

**Conclusion**

Despite the statistical findings throughout the two articles that were discussed, there are multiple articles that refute the idea of defense impacting a team’s win percentage. This could be because there are hidden issues in the initial regression analysis causing a Type I error, or there could be omitted variable bias possibly causing the total points variable to pick up the effect of the missing variable (Wheelan, 204). As discussed throughout the paper, multicollinearity could be present causing the analysis to be skewed (Wheelan, 2014). Exploring beyond the data is also another issue that could be present because the data being utilized is only from the 2019 NCAA football season. Including data from past seasons could change the statistical results altering the outcome. However, as it stands, the null hypothesis has been rejected because the independent variables were found to be statistically significant leading to the conclusion that defense does impact a team’s overall win percentage.

**Graphical user interface, diagram

Description automatically generatedAlteryx Screenshot**

**References**

Gallini, J. (2020). College football team stats 2019. <https://www.kaggle.com/jeffgallini/college->

football-team-stats-2019.

Pierson, B. (2018, January). "Defense Wins Championships" Proves True in Hoops. Retrieved September 18, 2020, from https://247sports.com/college/ucla/Article/Defense-Wins-Championships-Proves-True-in-College-Hoops-and-at-UCLA-113782988/

The Stats Zone. Does Defence Win Championships? (2016).

<https://www.thestatszone.com/archive/does-defence-win-championships-14048>.

Wheelan, C. J. (2014). Naked statistics: Stripping the dread from the data. New York: W.W.

Norton.