Investigating the Association between the Dallas Cowboys Offensive and Defensive Season Rankings and Margin of Victory

Rithvik Saravanan
rcs3348, rsaravanan@utexas.edu

Abstract: The purpose of this paper is to explore the best way to understand previous season ranking data and utilize it to predict the margin of victory for the Dallas Cowboys football team of the NFL in future seasons. To understand how the Dallas Cowboys' offensive and defensive rankings affect their margin of victory in games, we analyze a dataset that includes various season statistics for the Dallas Cowboys since 1960. These statistics are calculated and measured out of the full 16 games that the team plays each season. In this analysis, we are specifically interested in the Dallas Cowboys season rank for points scored, points allowed, yards gained, and yards allowed across each season in the dataset. In this study, we predict the expected margin of victory for specific rankings of the Dallas Cowboys on offense and defense. Additionally, we explore relationships between offense-related as well as defense-related predictors.

I. Background and Significance

The NFL ranks each team over the course of the season in numerous categories including offensive, defensive, and holistic statistics. The primary motivation for recording these statistics is to be able to track a team's progress over several years as well as compare teams to one another. These comparisons can help team owners, general managers, coaches, and other people in the organization's leadership better understand how to improve the team and its gameplay. Ultimately, the goal for each team is to show improvement from previous seasons and work their way toward the Super Bowl. These statistical analyses are typically the most common quantitative measure of understanding a team's strengths and weaknesses. In this study, our primary goal was to build a regression model that is practically understandable in the variables that it incorporates in order to reasonably predict a key significant variable for team leadership to utilize in their evaluation of team performance.

II. Methods

Data & Variables. All football-related statistics are stored and tracked by Pro Football Reference, the complete source for current and historical NFL, AFL, and AAFC players, teams, scores and leaders [2]. In this study, the statistics of focus include the rankings for the average number of points scored, the average number of yards gained, the average number of points allowed, the average number of yards allowed, and the average margin of victory over the course of a season (Appendix A). Points scored and yards gained are both offensive statistics while points allowed and yards allowed are both defensive statistics. This distinction allows us to explore whether associations exist between each side of the team.

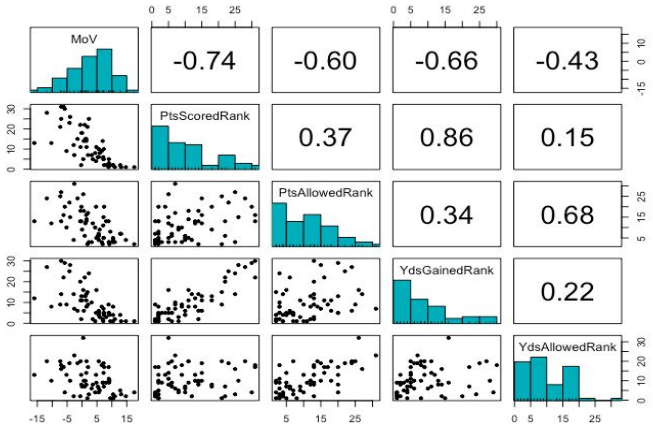
The average margin of victory can be a positive or negative value and is calculated by summing the score margins of a team's games and dividing by the total number of games played [4]. The significance of the margin of victory is that it indicates the level of dominance of a team for each game and over the course of the season. A higher margin of victory indicates that a team won by a higher number of points on average during the season and that they were more dominant in their matchups. This dominance can further be attributed to the strength of schedule or true offensive and defensive mismatches.

In this study, we will analyze these statistics for the Dallas Cowboys of the NFL for two particular reasons. First, the Dallas Cowboys were one of the original teams in the NFL and, as such, have accumulated season ranking data for approximately 60 years (since the 1960 season). This abundance of data provides for a better opportunity to more accurately build the regression model. Secondly, the season statistics for the Dallas Cowboys cover a wide range of values over the several decades of season. Some seasons include high playoff seedings and Super Bowl victories while other seasons include bottom five rankings in several of the categories. This wide range of data further improves our ability to fit a more accurate model to predict the average margin of victory.

Exploratory Data Analysis. **Fig 1.** Plotting the scatter plot with correlations for these specific variables, we notice that all of the predictors show a negative, linear, moderate to strong correlation with the response (MoV). We also observe that the rankings for points scored and yards gained as well as the rankings for the points allowed and yards allowed have positive, linear, strong correlations. This indicates that interaction effects between these pairs of predictors will be useful in improving the model. We also understand that logarithmic

transformations are not feasible for this model because the response variable (MoV) can be negative and this analysis is focused on understanding the prediction model rather than simply building the best prediction model [5].

III. Results



Since we sought to build the most practically understandable model to predict the average margin of victory, we first fit a base multiple linear regression model to predict average margin of victory from offensive and defensive season rankings in yards and points.

We refined this model by using the results of both forward and backward stepwise selection as well as best subsets regression. Both forward and backward stepwise selection resulted in using only the variables of rankings for points scored, points allowed, and yards allowed. Best subsets regression also showed that these same three predictors resulted in the most optimal model with a low SSE, high adjusted R^2, low Mallow's C_p, and a high BIC (Appendix B and C).

To investigate the relationship between the offensive and defensive predictors, we plotted them and checked their correlation coefficients. The plots are shown in **Fig 2** and **Fig 3**.

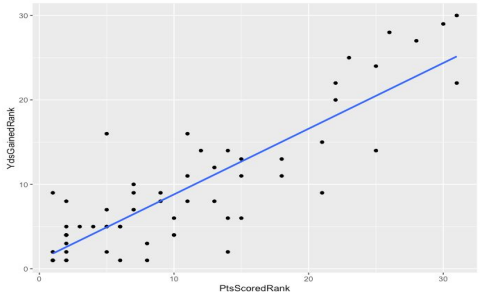


Fig 2. Scatter plot of offensive predictors

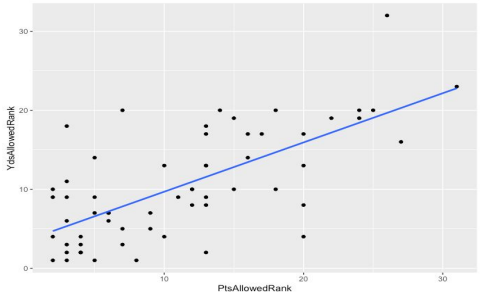


Fig 3. Scatter plot of defensive predictors

The correlation for the offensive predictors is 0.86 and the correlation for the defensive predictors is 0.68. Since the ranking for yards gained was not included in the model from the regression selection techniques, we incorporated only the interaction effect between the defensive predictors of rankings for average points allowed and average yards allowed.

To verify the validity of this model, we checked the VIF values and noticed that there was high multicollinearity between the predictors because the VIF values were approximately 5. To account for this, we centered the predictors by shifting them to fall around the mean value. By doing so, we reduced the VIF values to approximately 1 and mitigated the effect of multicollinearity. The final model we used for our predictions is represented by the equation

$$y = 2.6351 - 0.4595x_1 - 0.2832x_2 - 0.1773x_3 + 0.0097x_2x_3$$

where y represents the average margin of victory, x_1 represents the season rank in average points scored, x_2 represents the season rank in average points allowed, and x_3 represents the season rank in average yards allowed. This model has an F-value of 33.5 with a p-value of

4.778×10^{-14}. Since this p-value is significantly less than the significance threshold of 0.05, our final model is very effective. This model also resulted in an R^2 of 0.7128 and an adjusted R^2 of 0.6915. See Appendix D, E, F, and G for the summary and outputs of this regression model.

IV. Conclusions

As expected, the offensive statistics of rankings for average points scored and average yards gained as well as the defensive statistics of rankings for average points allowed and average yards allowed were highly correlated. These findings were reasonable because teams that gain more yards are more likely to possess the ball in the red zone (within the 20 yard line) and will accordingly have more opportunities to score touchdowns and field goals to add to the overall point total [3]. Similarly, teams that allow opponents to gain more yards are more likely to give them more opportunities to score. We utilized this knowledge to incorporate an appropriate interaction effect in our final regression model.

Since we sought to build the most practically understandable model to predict the average margin of victory, the model coefficients are also of interest because they communicate that higher rankings in any of the three predictors result in a decrease in the average margin of victory. This is also reasonable because a team's offense or defense that is ranked higher in any of the categories is generally considered to be inferior to a team's offense or defense that is ranked lower in that same category. Specifically, the coefficient for the most important predictor (ranking for average points scored) in the model indicates that, holding the other predictors constant, an increase in 1 ranking position in average points scored results in a predicted decrease in the average margin of victory by 0.4595 points. This is a notable insight because this allows the team's leadership to identify that offensive scoring is the single, most significant factor in determining margin of victory [1]. Since our model was built primarily for understandability, this takeaway effectively allows the team ownership to focus on re-signing key offensive players and acquiring more offensive talent through free agency and the NFL Draft.

The adjusted coefficient of determination of 0.6915 indicates that 69.15% of the variation in the average margin of victory is accounted for by the predictor variables. In other words, when these predictor variables are considered, the variation in average margin of victory is reduced by 69.15%. According to this regression model, for a season where the Dallas Cowboys are ranked first in offensive and defensive yards and points, their margin of victory is predicted to be 12.1116 points. Additionally, the model predicts with 95% confidence that the margin of victory would be between 3.7889 and 20.4342 points. For seasons where the Dallas Cowboys are ranked last in offensive and defensive yards and points, this model predicts their margin of victory to be -13.2229 points. Additionally, the model predicts with 95% confidence that the margin of victory would be between -20.9352 and -5.5107 points.

This regression model helps explain that average margin of victory is an important statistic that can convey how well or how poorly a team is playing throughout the season. It can be used by each team's leadership to identify weaknesses in the team's play and address these issues by acquiring free agent players and drafting specific position players in the NFL Draft. This study shows that the three most important predictors of average margin of victory are the rankings for average points scored, average points allowed, and average yards allowed (in that order).

References

1. “A Complete History of NFL Margins of Victory.” *ELDORADO*, www.eldo.co/despite-its-weird-scores-2015-had-the-most-normal-margins-of-victory-in-nfl-history.html.

2. “Dallas Cowboys Team Encyclopedia.” *Pro*, www.pro-football-reference.com/teams/dal/.

3. “Football Forecasting - Margin-of-Victory Model.” *EdsCave*, www.edscave.com/margin-of-victory-best-fit.html.

4. “NFL Team Average Scoring Margin.” *NFL Football Stats - NFL Team Average Scoring Margin on TeamRankings.com*, www.teamrankings.com/nfl/stat/average-scoring-margin.

5. Paine, Neil. “What's the Correct Margin-of-Victory Cap For the NFL?” *FootballPerspective.com*, 23 Dec. 2013, www.footballperspective.com/whats-the-correct-margin-of-victory-cap-for-the-nfl/.

Appendix

Appendix A

MoV	PtsScoredRank	PtsAllowedRank	YdsGainedRank	YdsAllowedRank
7.1	6	11	1	9
0.9	22	6	22	7
1.4	14	13	14	8
7.2	5	5	5	14
-6.2	31	16	22	17
7.2	5	15	7	19

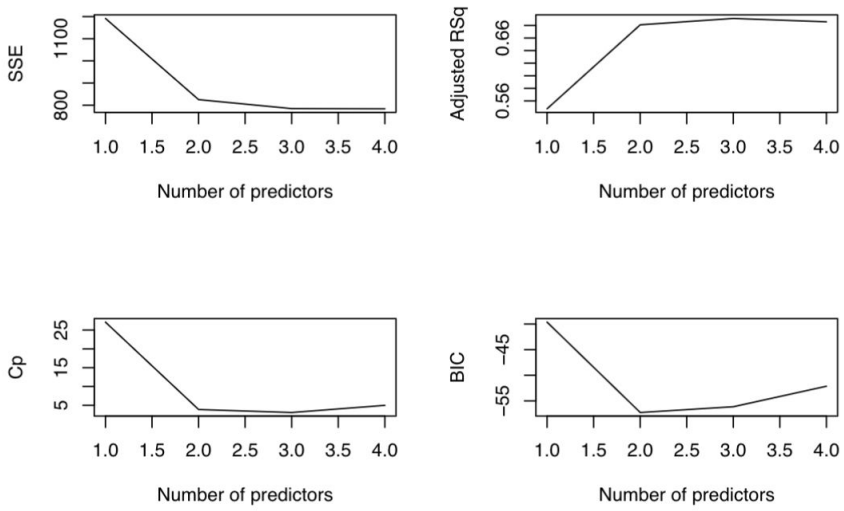
This table shows an example of what types of values are stored in each variable. Margin of victory (MoV) is a decimal number because it is an average over multiple games over the course of the season and can be negative or positive. The other four variables only contain integer values because they indicate the team's ranking for a season compared to the other teams in the NFL.

Appendix B

```
Call:
lm(formula = MoV ~ PtsScoredRank + PtsAllowedRank + YdsGainedRank +
    YdsAllowedRank, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-17.1024  -0.3633   0.5633   1.8565   6.6565

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    12.386533   1.079058  11.479 3.17e-16 ***
PtsScoredRank  -0.476896   0.119479  -3.991 0.000196 ***
PtsAllowedRank -0.242511   0.100176  -2.421 0.018812 *
YdsGainedRank  -0.004942   0.128988  -0.038 0.969575
YdsAllowedRank -0.144042   0.104504  -1.378 0.173678
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.942 on 55 degrees of freedom
Multiple R-squared:  0.6848, Adjusted R-squared:  0.6619
F-statistic: 29.88 on 4 and 55 DF,  p-value: 3.212e-13
```

This figure shows the summary output for the base multiple regression model using the first order of all four predictors. Notice that the R^2 is 0.6848 and the adjusted R^2 is 0.6619. The F-value for this model is 29.88 and the p-value for is 3.212×10^{-13}.

Appendix C



These plots show the outputs for SSE, adjusted R^2, Mallow's C_p, and BIC from best subsets regression. Notice that three predictors are the optimal number to use in the model because it results in the smallest SSE, highest R^2, smallest Mallow's C_p, and smallest BIC. The three predictors mentioned in this model selection method are the rankings for average points scored, average points allowed, and average yards allowed.

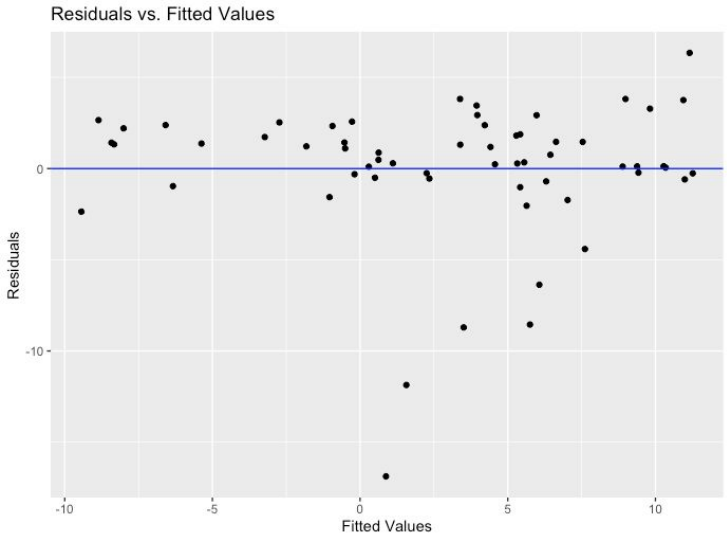
Appendix D

```
Call:
lm(formula = MoV ~ PtsScoredRank.c + PtsAllowedRank.c + YdsAllowedRank.c +
    PtsAllowedRank.c * YdsAllowedRank.c, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-16.8454  -0.8006   0.8375   1.7857   5.8606

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      2.635055   0.588456   4.478 3.94e-05 ***
PtsScoredRank.c                 -0.459528   0.062654  -7.334 1.18e-09 ***
PtsAllowedRank.c                -0.283174   0.100524  -2.817  0.00676 **
YdsAllowedRank.c                -0.177300   0.099274  -1.786  0.07972 .
PtsAllowedRank.c:YdsAllowedRank.c  0.009743   0.009383   1.038  0.30373
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.775 on 54 degrees of freedom
Multiple R-squared:  0.7128, Adjusted R-squared:  0.6915
F-statistic:  33.5 on 4 and 54 DF,  p-value: 4.778e-14
```

This figure shows the summary output for the final multiple regression model using the three centered predictors recommended by the model selection methods and the interaction effect between rankings for average points allowed and average yards allowed. Notice that the R^2 is

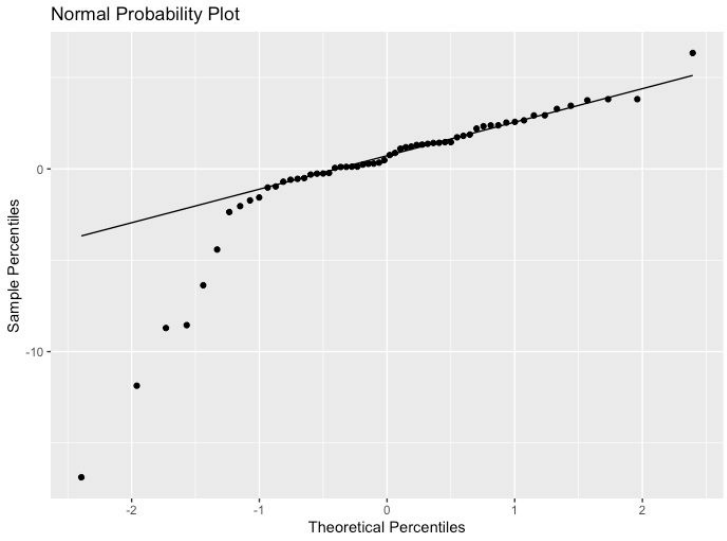
0.7128 and the adjusted R^2 is 0.6915. The F-value for this model is 33.5 and the p-value for is 4.778×10^{-14}. All of these criteria show notable improvements over the baseline regression model.

Appendix E



This figure shows the residual plot for the final regression model. From this residual plot, we observe that there is approximately equal variance because the residuals are approximately equal in average magnitude for the fitted values across the plot. Excluding the few outliers near the bottom of the plot, we note that there is no obvious pattern since the fitted values are scattered mostly randomly across the plot. Therefore, this plot indicates linearity.

Appendix F



This figure shows the normal probability plot for the final regression model. From this normal probability plot, we can observe that the data points form an approximately straight line and line up mostly along the line shown in the plot. This indicates that the normal distribution is a good model because the plot shows no significant deviation from a normal distribution of error terms. Since there is deviation from the line around the extremes (specifically at the lower extreme), this plot indicates heavy tails. Otherwise, the assumption that the residuals are normally distributed is approximately met.

Appendix G

```
 PtsScoredRank.c                  PtsAllowedRank.c
        1.241950                          2.348999
YdsAllowedRank.c PtsAllowedRank.c:YdsAllowedRank.c
        1.897890                          1.312197
```

This figure shows the VIF values for the final regression model with centered predictors. Notice that all VIF values are less than 2.5. This indicates that there is very little multicollinearity in the model and shows significant improvements over the baseline regression model and the model without centered predictors.