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

Rithvik Saravanan

December 7, 2020

Abstract

To explore how the ranking of the Dallas Cowboys on offense and defense affect their margin of victory, we explore a dataset that includes various season statistics for the Dallas Cowboys football team of the NFL. 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 measured on a scale of 1 to 32 (representing the 32 teams in the NFL). 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.

Background and Significance

The NFL ranks each team over the course of the season in various categories including offensive and defensive statistics. These rankings can help identify the margin of victory for a specifically-ranked team and identify their level of dominance in game matchups where both teams have specific offensive and defensive rankings in yards and points. In this study, I will particularly focus on the Dallas Cowboys.

`PtsScoredRank` indicates the season rank in points scored and `YdsGainedRank` indicates the season rank in yards gained. `PtsAllowedRank` indicates the season rank in points allowed and `YdsAllowedRank` indicates the season rank in yards allowed. `PtsScoredRank` and `YdsGainedRank` are both primarily offensive statistics and `PtsAllowedRank` and `YdsAllowedRank` are both primarily defensive statistics. These rankings are accurate measures of how the team performed over a typical season relative to the other 31 NFL teams.

In this analysis, using these predictor variables, we analyze the relationship between offensive and defensive rankings and `MoV`, the Dallas Cowboys average margin of victory over a season. `MoV` 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.

Methods

To build the most apt model to answer these research topics, I fit a multiple linear regression model where the predictors of offensive and defensive season rankings in yards and points are used to predict the margin of victory. I built this model by comparing the results and fit of both forward and backward stepwise selection as well as best subsets regression. I also investigated whether there are relationships between any of the predictors, and if so, I considered adding interaction effects to the model. To account for multicollinearity in the model, I centered each of the predictors.

The resulting model from these model selection criteria is shown below.

```
# load data
data <- read.csv('./dallas_cowboys_season_data.csv')
mydata <- data[c("MoV", "PtsScoredRank", "PtsAllowedRank",
                 "YdsGainedRank", "YdsAllowedRank")]

# center the predictors
mydata <- mydata %>%
          mutate(PtsScoredRank.c = PtsScoredRank - mean(PtsScoredRank),
                 PtsAllowedRank.c = PtsAllowedRank - mean(PtsAllowedRank),
                 YdsGainedRank.c = YdsGainedRank - mean(YdsGainedRank),
                 YdsAllowedRank.c = YdsAllowedRank - mean(YdsAllowedRank))

# fit the regression model with centered predictors
reg <- lm(MoV ~ PtsScoredRank.c + PtsAllowedRank.c + YdsAllowedRank.c +
                PtsAllowedRank.c * YdsAllowedRank.c, mydata)

summary(reg)
```

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

```
##      Min       1Q   Median       3Q      Max
## -16.8757  -0.5145   0.6177   1.9602   6.3420
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      2.655825   0.608736   4.363 5.69e-05 ***
## PtsScoredRank.c                 -0.472129   0.064935  -7.271 1.36e-09 ***
## PtsAllowedRank.c                -0.263466   0.104595  -2.519   0.0147 *
## YdsAllowedRank.c                -0.152154   0.102751  -1.481   0.1444
## PtsAllowedRank.c:YdsAllowedRank.c  0.006267   0.009638   0.650   0.5182
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.927 on 55 degrees of freedom
## Multiple R-squared:  0.6872, Adjusted R-squared:  0.6645
## F-statistic: 30.21 on 4 and 55 DF,  p-value: 2.613e-13
```

The diagnostics for this model are included in the appendix.

Results

```
# create a new observation to predict MoV for ranking first in all 4 categories
rankedFirst <- data.frame(PtsScoredRank.c = 1 - mean(mydata$PtsScoredRank),
                          PtsAllowedRank.c = 1 - mean(mydata$PtsAllowedRank),
                          YdsGainedRank.c = 1 - mean(mydata$YdsGainedRank),
                          YdsAllowedRank.c = 1 - mean(mydata$YdsAllowedRank))

# calculate the corresponding prediction
predict(reg, rankedFirst, interval = "prediction")
```

```
##        fit      lwr     upr
## 1 12.11155 3.788905 20.4342
```

```
# create a new observation to predict mean MoV for ranking last in all 4 categories
rankedLast <- data.frame(PtsScoredRank.c = 32 - mean(mydata$PtsScoredRank),
                         PtsAllowedRank.c = 32 - mean(mydata$PtsAllowedRank),
                         YdsGainedRank.c = 32 - mean(mydata$YdsGainedRank),
                         YdsAllowedRank.c = 32 - mean(mydata$YdsAllowedRank))

# calculate the corresponding prediction
predict(reg, rankedLast, interval = "confidence")
```

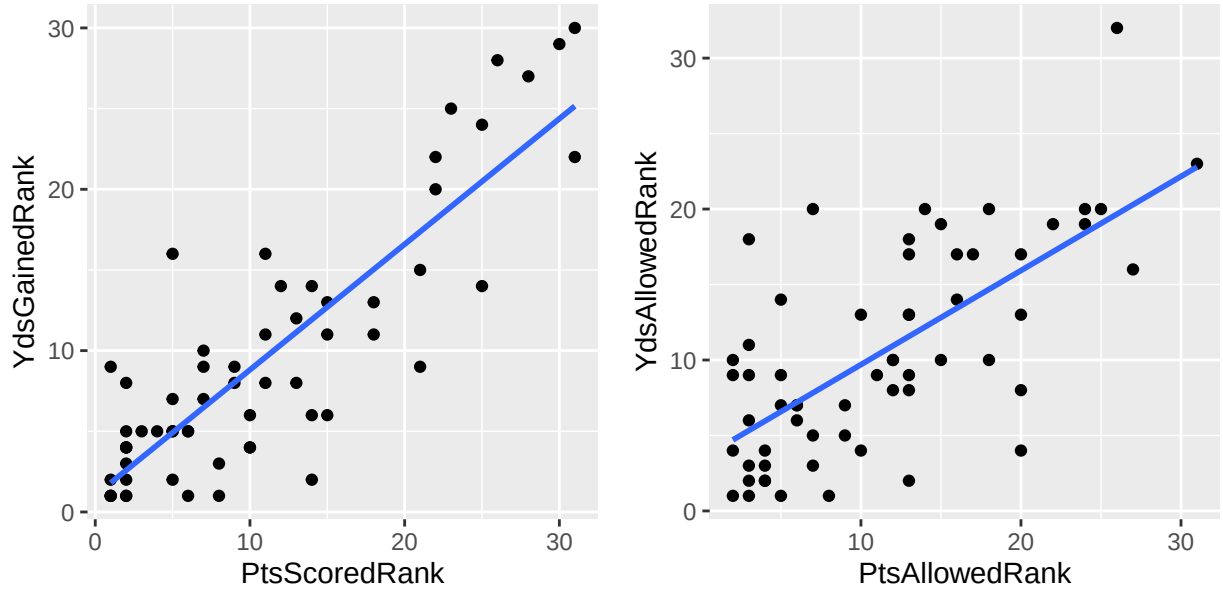
```
##         fit       lwr       upr
## 1 -13.22294 -20.93517 -5.510697
```

```
# relationship and correlation between points scored and yards gained
ggplot(mydata, aes(x = PtsScoredRank, y = YdsGainedRank)) +
  geom_point() +
  geom_smooth(method = lm, se = FALSE)

cor(mydata$PtsScoredRank, mydata$YdsGainedRank)
```

```
## [1] 0.8586953
```

```
# relationship and correlation between points allowed and yards allowed
ggplot(mydata, aes(x = PtsAllowedRank, y = YdsAllowedRank)) +
  geom_point() +
  geom_smooth(method = lm, se = FALSE)

cor(mydata$PtsAllowedRank, mydata$YdsAllowedRank)
```

```
## [1] 0.6814417
```



Conclusions

From this model, we can reach conclusions for the research topics posed earlier.

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 by this regression model. Additionally, the model predicts with 95% confidence that the margin of victory would be between 3.7889 and 20.4342.

For seasons where the Dallas Cowboys are ranked last in offensive and defensive yards and points, their margin of victory is predicted to be -13.2229 by this regression model. Additionally, the model predicts with 95% confidence that the margin of victory would be between -20.9352 and -5.5107.

The correlation between offensive points scored rank and yards gained rank is 0.8587. This indicates a strong, positive, linear relationship between the two predictors. From the scatter plot, we notice that the data points fall relatively along the line in the plot and that there is indeed a relationship between offensive points gained rank and yards gained rank. Therefore, the interaction effect can be useful when using these two predictors in the model. However, since we removed offensive yards gained from the model due to the model selecion criteria, this interaction effect was not included.

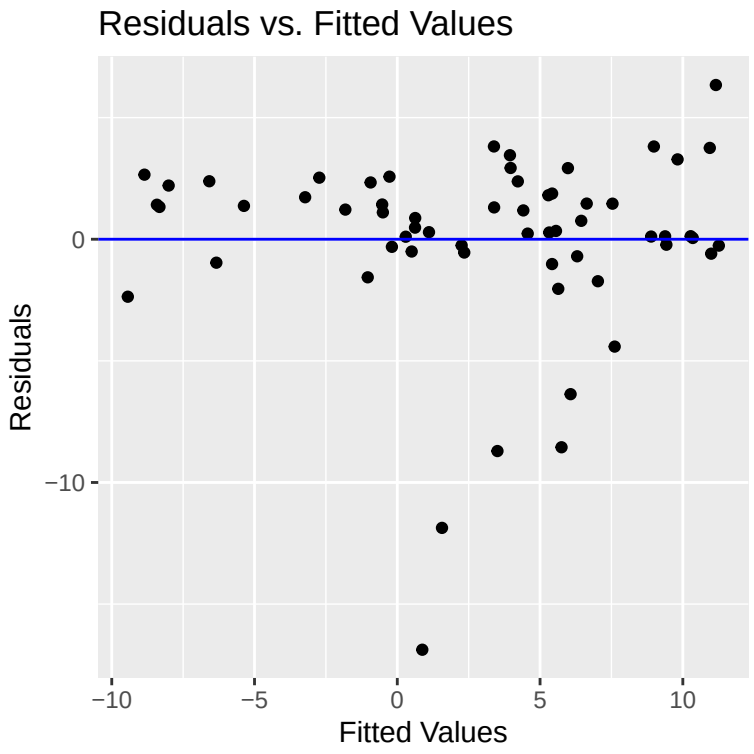
The correlation between defensive points allowed rank and yards allowed rank is 0.6814. This indicates a moderate, positive, linear relationship between the two predictors. From the scatter plot, we notice that the data points fall relatively along the line in the plot and that there is indeed a relationship between defensive

points allowed rank and yards allowed rank. Therefore, the interaction effect can be useful when using these two predictors in the model. Since both of these predictors were included in the model, this interaction term was incorporated as well.

Appendix

Residual plot of the regression model (diagnostic #1)

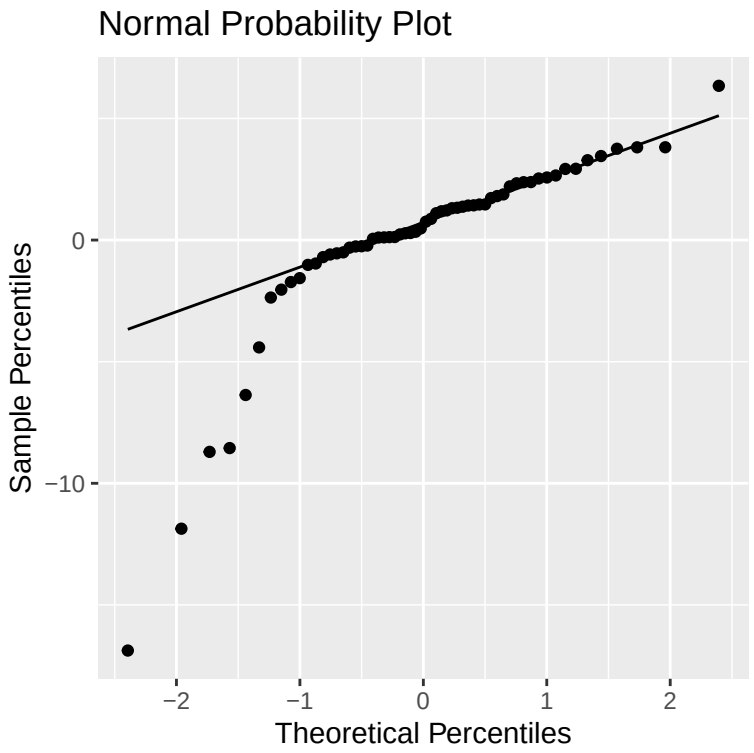
```
# residuals vs. fitted values
mydata$resids <- residuals(reg)
mydata$predicted <- predict(reg)
ggplot(mydata, aes(x = predicted, y = resids)) +
       geom_point() +
       geom_hline(yintercept = 0, color = "blue") +
       labs(title ="Residuals vs. Fitted Values",
            x = "Fitted Values",
            y = "Residuals")
```



Normal probability plot of the regression model (diagnostic #2)

```
# normal probability plot
ggplot(mydata, aes(sample = resids)) +
       stat_qq() +
       stat_qq_line() +
```

```
       labs(title = "Normal Probability Plot",
            x = "Theoretical Percentiles",
            y = "Sample Percentiles")
```



VIF for the regression model (multicollinearity)

```
# check for multicollinearity
vif(reg)
```

```
##                  PtsScoredRank.c                  PtsAllowedRank.c
##                         1.233146                         2.388315
##                YdsAllowedRank.c PtsAllowedRank.c:YdsAllowedRank.c
##                         1.929066                         1.318535
```

References

(1) Pro Football Reference, Dallas Cowboys, NFL (https://www.pro-football-reference.com/teams/dal/)

(2) TeamRankings, NFL Team Average Scoring Margin (https://www.teamrankings.com/nfl/stat/average-scoring-margin)