410FinalProject

Ross Bechtel

5/3/2021

Project Statement

The question I will aim to answer in this project is "What linear relationships do oddsmakers use to create money line odds for NFL games?".

For those unfamiliar with sports betting and sports odds, money line odds are the odds used to represent the probability a team has to win a given game. They are strangely formatted but can be converted to an implied probability. The favorite to win the game has a negative statistic while the underdog has a positive statistic. If a team is listed as -250 to win a game, a bettor would have to risk \$250 to profit \$100. If a team is listed as +250 to win a game, a bettor would have to risk \$100 to profit \$250. In this scenario, the implied probability for the favorite is .7143 and the implied probability for the underdog is .2857. For favorites, the conversion takes place using the following formula:

$$\frac{-(MLOdds)}{-(MLODDS)+100} = \frac{-(-250)}{-(-250)+100} = \frac{250}{250+100} = 250/350 = .7143$$

For underdogs, the conversion takes place using a similar formula:

$$\frac{100}{MLODDS+100} = \frac{100}{250+100} = \frac{100}{350} = .2857$$

One thing to note is that in a head to head matchup, the sum of the implied probabilities for two opposing sides will not always sum to 1. In most cases, they will sum to more than 1. This is because the oddsmakers and bookies (those who make transactions with bettors) manipulate the odds so that they always win. In giving both sides better odds than they actually have, they guarantee a lower payout for gamblers and a profit for themselves.

My interest in this question and this topic comes from my interest in the use of statistics in sports. Oddsmakers are the best in the world at predicting the outcome of sports events (if they weren't, they would be out of business). To generate their odds, they use advanced models and algorithms that are a mystery to the general public. In this project, I hope to gain some insight into what variables are important to oddsmakers when generating their NFL odds.

Data Description

The first dataset contains data from the 2020-2021 NFL season and the money line odds associated with each team for each game from that season. This dataset is important because it contains the ML odds and potential variables that could impact those odds. This data comes from https://www.sportsbookreviewsonline.com/scoresoddsarchives/nfl/nfloddsarchives.htm

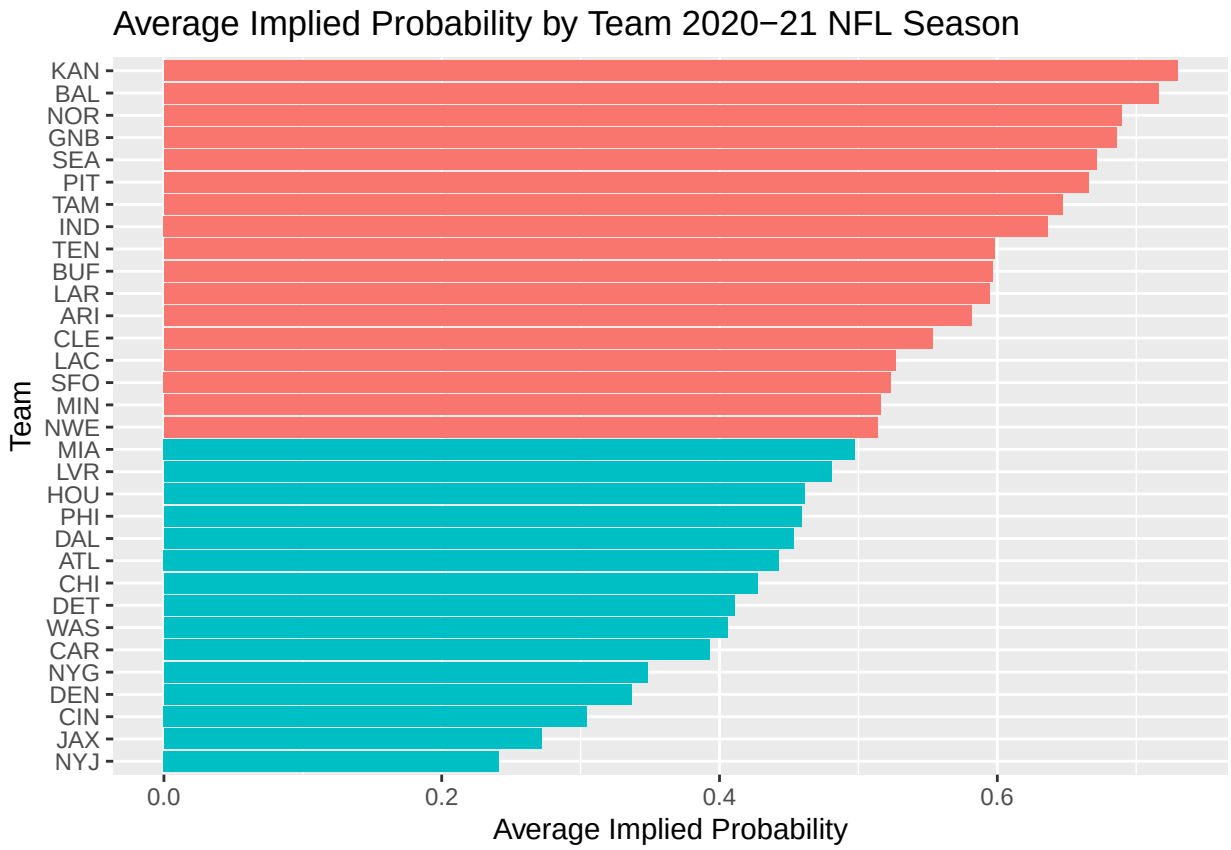
The second dataset consists of weekly data on each NFL team from the 2020-2021 NFL season. It contains variables like quarterback rating for the team's QB, the team's offensive yards per play, defensive yards per play, aturnover margin, point differential among others. All of these statistics are cumulative up to each game week. This dataset is important because it contains many variables that could impact ML odds. This data comes from https://www.pro-football-reference.com 's query tool.

Exploratory Data Analysis

Note: Code including the reading of data, subsetting of data, and cleaning of data is not included in this pdf out of reading convenience becasue it was over 150 lines and very repetitive. Please check rmd file for this code.

Let's first explore the odds data:

```
# Create column for favorite or underdog
odds$wasFav <- ifelse(odds$ML < 0, 1, 0)
# Calculate implied odds
odds$implied <- ifelse(odds$wasFav == 1,
                       -odds$ML / (-odds$ML + 100),
                       100 / (odds$ML + 100))


# Get average implied odds by team
avgs <- odds %>%
  group_by(Team) %>%
  summarise(avg_implied = mean(implied))
avgs$color <- ifelse(avgs$avg_implied >= 0.5, 'blue', 'red')


ggplot(avgs, aes(reorder(Team, avg_implied), avg_implied, fill=color)) +
  geom_col() +
  coord_flip() +
  labs(y="Average Implied Probability", x="Team") +
  ggtitle("Average Implied Probability by Team 2020-21 NFL Season") +
  theme(legend.position = 'none')
```

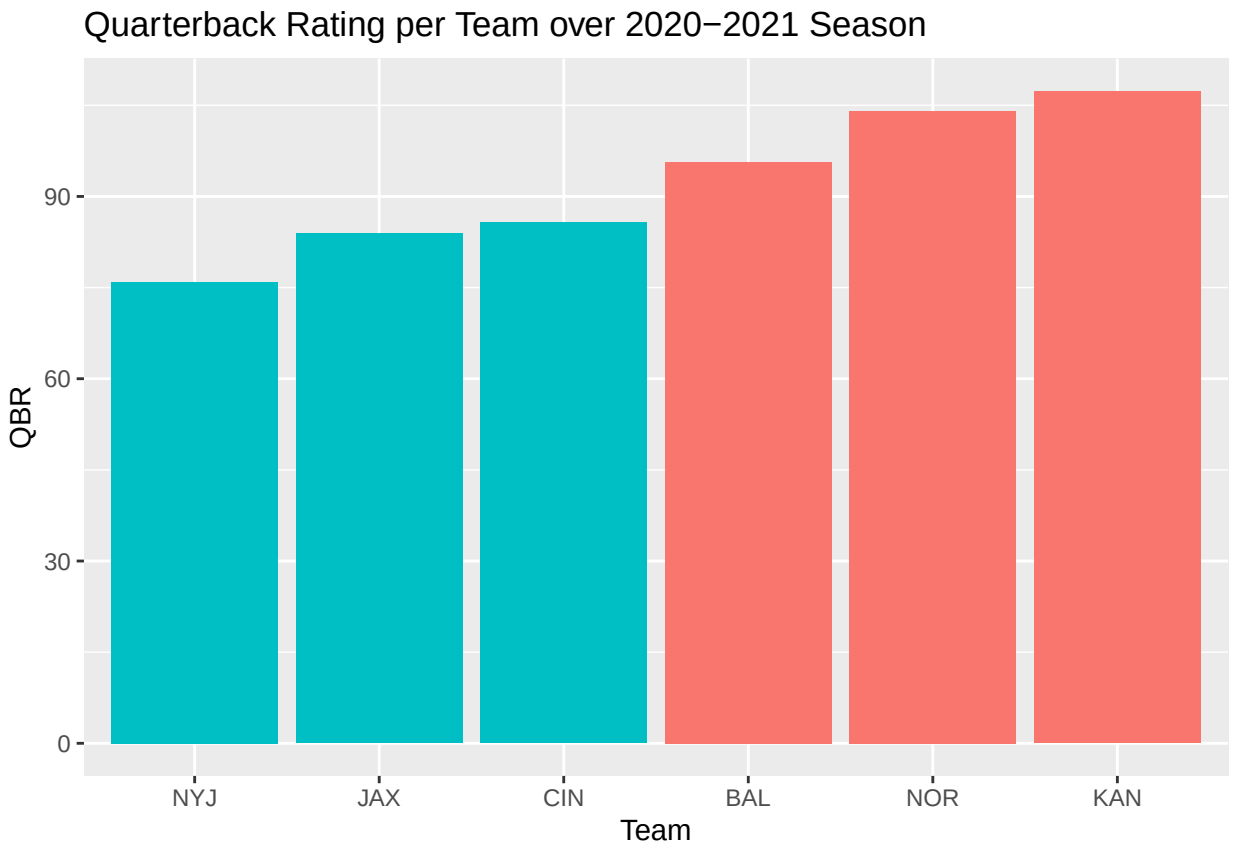


This plot shows that the oddsmakers tended to favor the Kansas City Chiefs, Baltimore Ravens, and New Orleans Saints the most. It also shows that the oddsmakers tended to believe the least in the New York Jets, Jacksonville Jaguars, and Cincinnati Bengals. So, what did the Chiefs, Ravens and Saints do right? And, what did the Jets, Jaguars and Bengals do wrong?

Let's take a look at the end of the season stats for these teams:

```
# Get best and worst teams oonly
bestAndWorst <- subset(week17,
                       week17$Tm %in% c('KAN','BAL','NOR','CIN','JAX','NYJ'))
bestAndWorst$color <- c('Best','Best','Best','Worst','Worst','Worst')


ggplot(bestAndWorst, aes(reorder(Tm, Rate), Rate, fill=color)) +
  geom_col() +
```

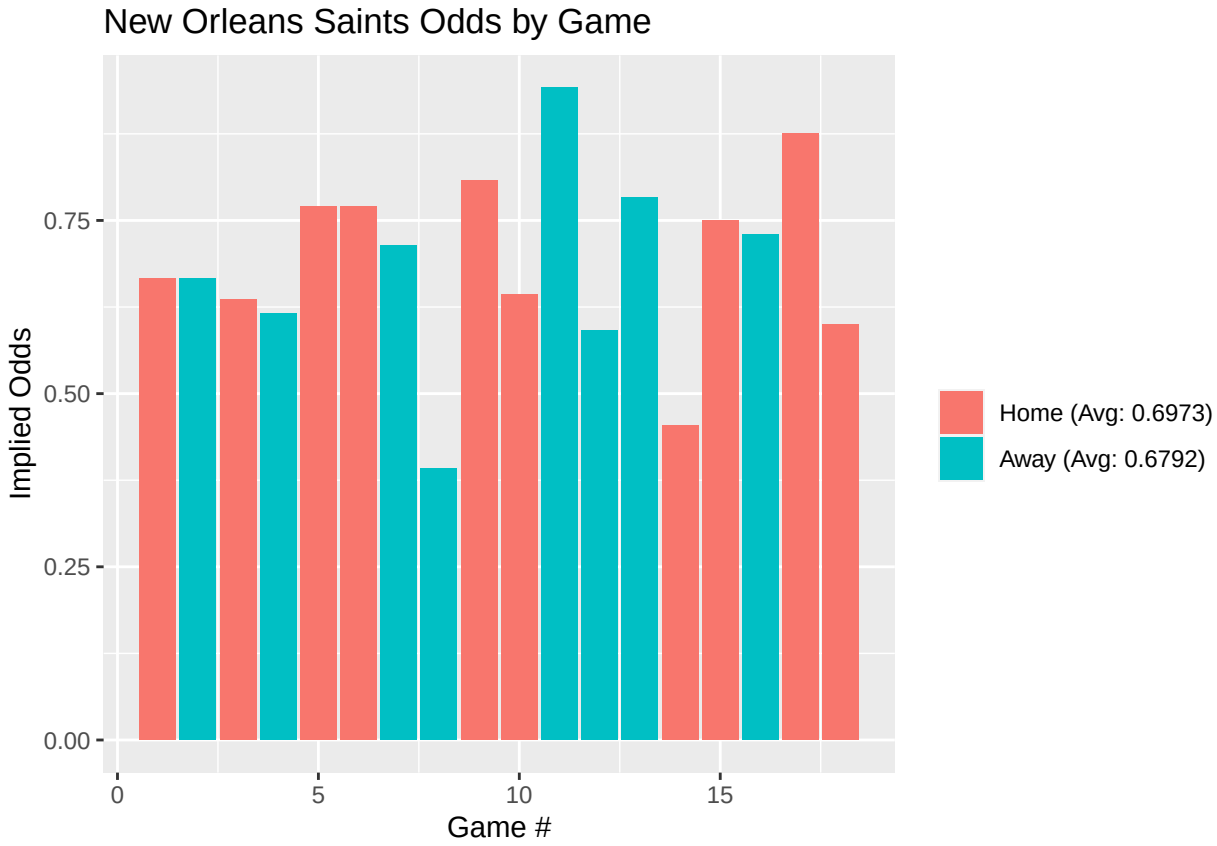
```
  labs(x="Team",y="QBR") +
  theme(legend.position = 'none') +
  ggtitle("Quarterback Rating per Team over 2020-2021 Season")
```



Quarterback rating (QBR) is a measure of how good a quarterback is. It is an all inclusive stat that accounts for traditional quarterback stats like touchdowns, completion %, total yards, and interceptions. The QB is the most important player on any NFL team and makes the most difference out of any other player. So, it would make sense that a team with a good QBR would have better odds than a team with a low QBR. This plot definitely seems to agree with that, showing how the top three favored teams all had higher QBRs than the bottom three.

Returning to the odds data, let's look at weekly changes in odds for a single team - The New Orleans Saints (my favorite team):

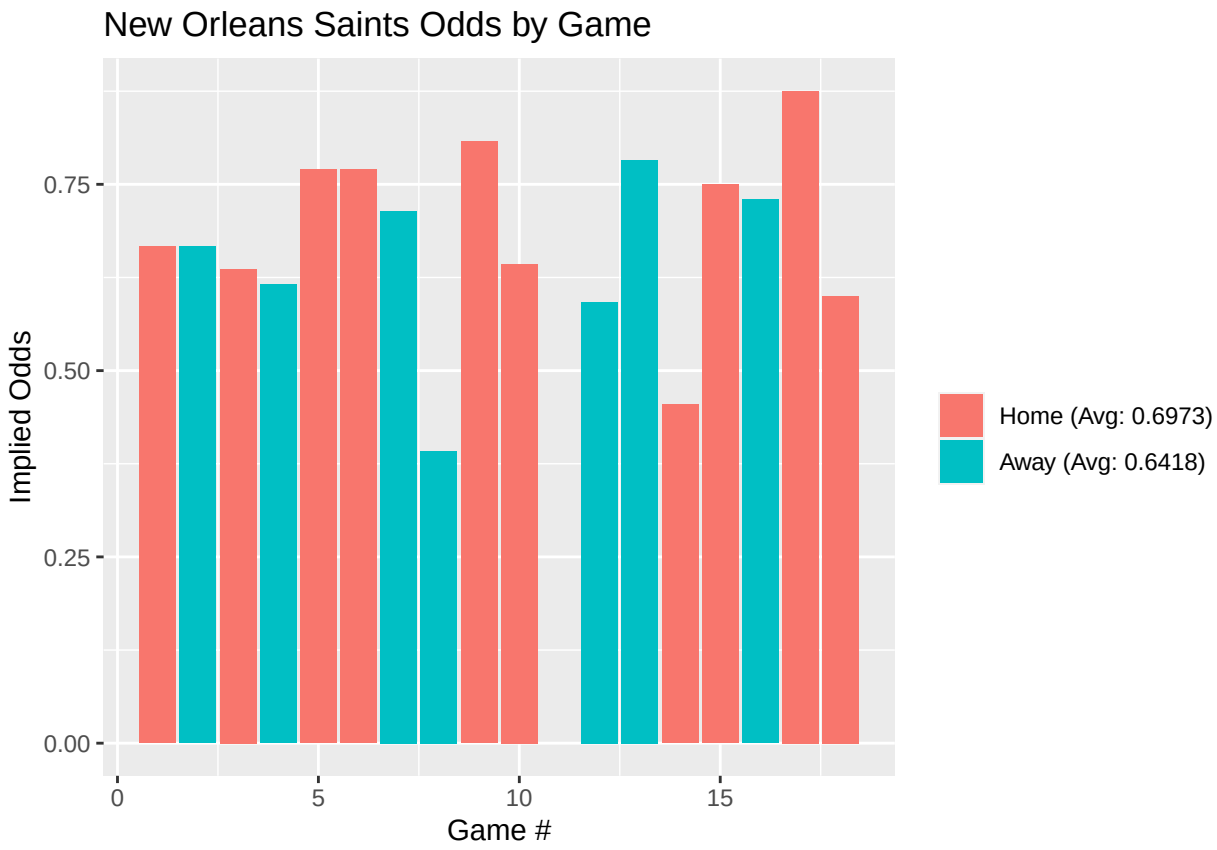
```r
# Subset for Saints and summarise
saintsOdds <- subset(odds, odds$Team == 'NOR')
saintsOdds$game <- seq(1:length(saintsOdds$Date))
saintsOddsSum <- saintsOdds %>%
  group_by(VH) %>%
  summarise(odds=mean(implied))


ggplot(saintsOdds, aes(game, implied, fill=VH)) +
  geom_col() +
  labs(x="Game #", y="Implied Odds") +
  ggtitle("New Orleans Saints Odds by Game") +
  scale_fill_discrete(name="",
                      labels=c(paste0("Home (Avg: ",
                                      round(saintsOddsSum[1,2],
                                            4),")"),
                               paste0("Away (Avg: ",
                                      round(saintsOddsSum[2,2],
                                            4),")")))
```



Home field advantage is a common effect that many believe exist in all sports. What do the oddsmakers have to say about this? In the 2020-2021 season, the Saints were given more favorable odds than when they played at home than they did when they played on the road. The difference in the average implied odds is about .02. This value is also skewed due to the fact that when the Saints visited the Denver Broncos in their 11th game of the season, the Broncos were without their top 3 QB options. Removing that outlier yields the following plot:

```
# Remove week 11 and make same plot
saintsOddsSum <- saintsOdds[-11,] %>%
  group_by(VH) %>%
  summarise(odds=mean(implied))
ggplot(saintsOdds[-11,], aes(game, implied, fill=VH)) +
  geom_col() +
```

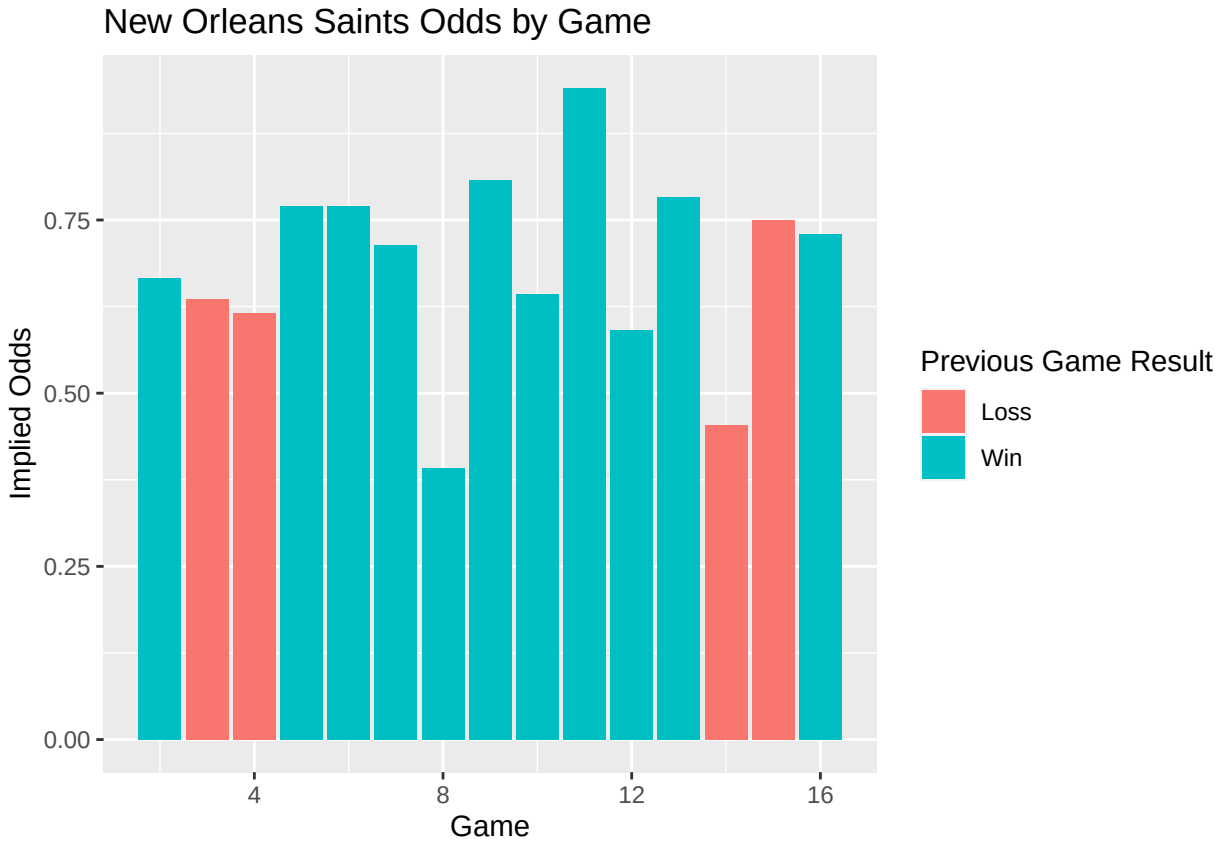
```
  labs(x="Game #", y="Implied Odds") +
  ggtitle("New Orleans Saints Odds by Game") +
  scale_fill_discrete(name="",
                      labels=c(paste0("Home (Avg: ",
                                      round(saintsOddsSum[1,2],
                                            4),")"),
                               paste0("Away (Avg: ",
                                      round(saintsOddsSum[2,2],
                                            4),")")))
```



Now, the difference in average implied odds is about 0.055, which seems significant.

```
# Remove week 1 as their is no previous week and remove playoffs
saintsOddsReg <- saintsOdds[-c(1,17,18),]
saintsOddsReg$prevWeek <- c('W','L','L','W','W','W','W','W',
```

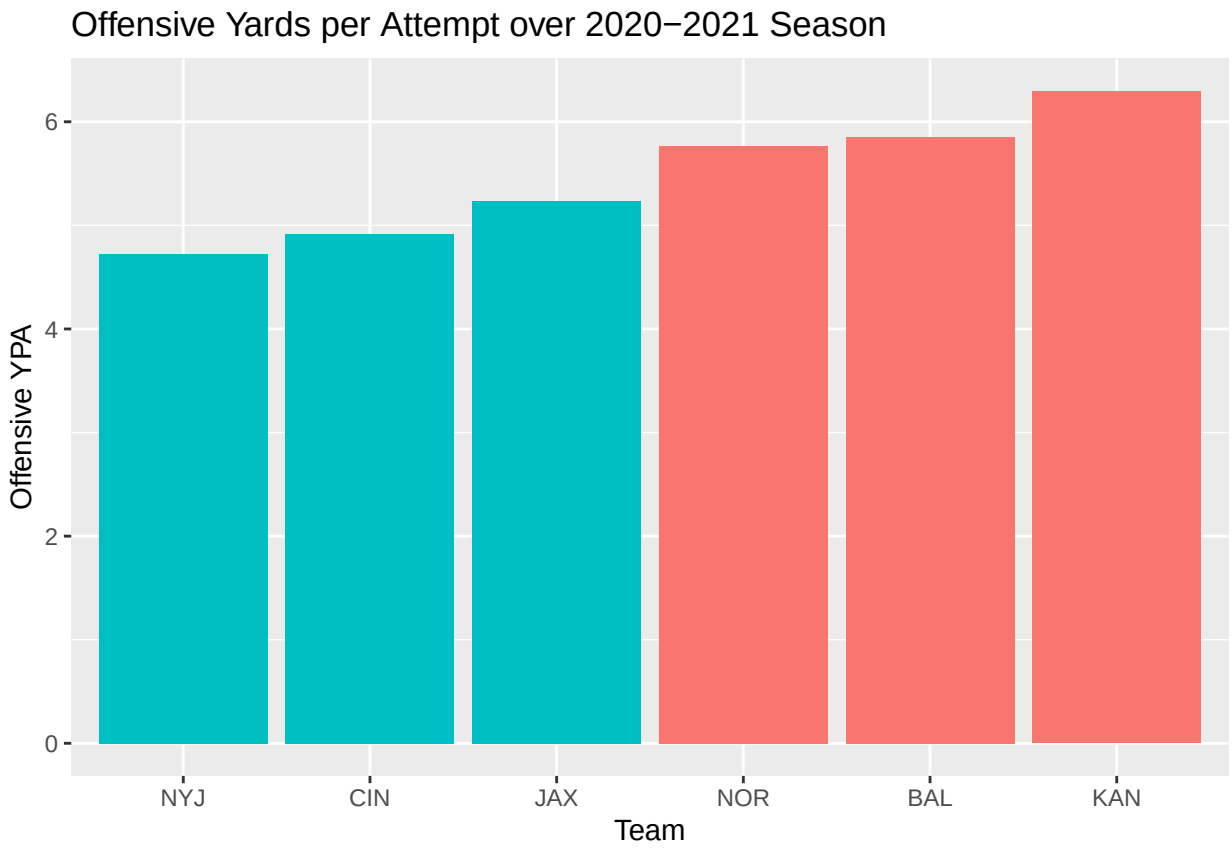
```
                          'W','W','W','W','L','L','W')
ggplot(saintsOddsReg, aes(game, implied, fill=prevWeek)) +
  geom_col() +
  labs(x="Game",y="Implied Odds") +
  ggtitle("New Orleans Saints Odds by Game") +
  scale_fill_discrete(name="Previous Game Result",labels=c("Loss","Win"))
```



The ability of a football team changes throughout a given season. For this reason, recent results are important in determining how good a team is at a certain point in time. The above plot is colored by the previous game's result. The Saints saw a decrease in odds in 3/4 of the games they played following a loss. Although the Saints did not lose many games, this seems to suggest that oddsmakers do care about what a team has done recently.

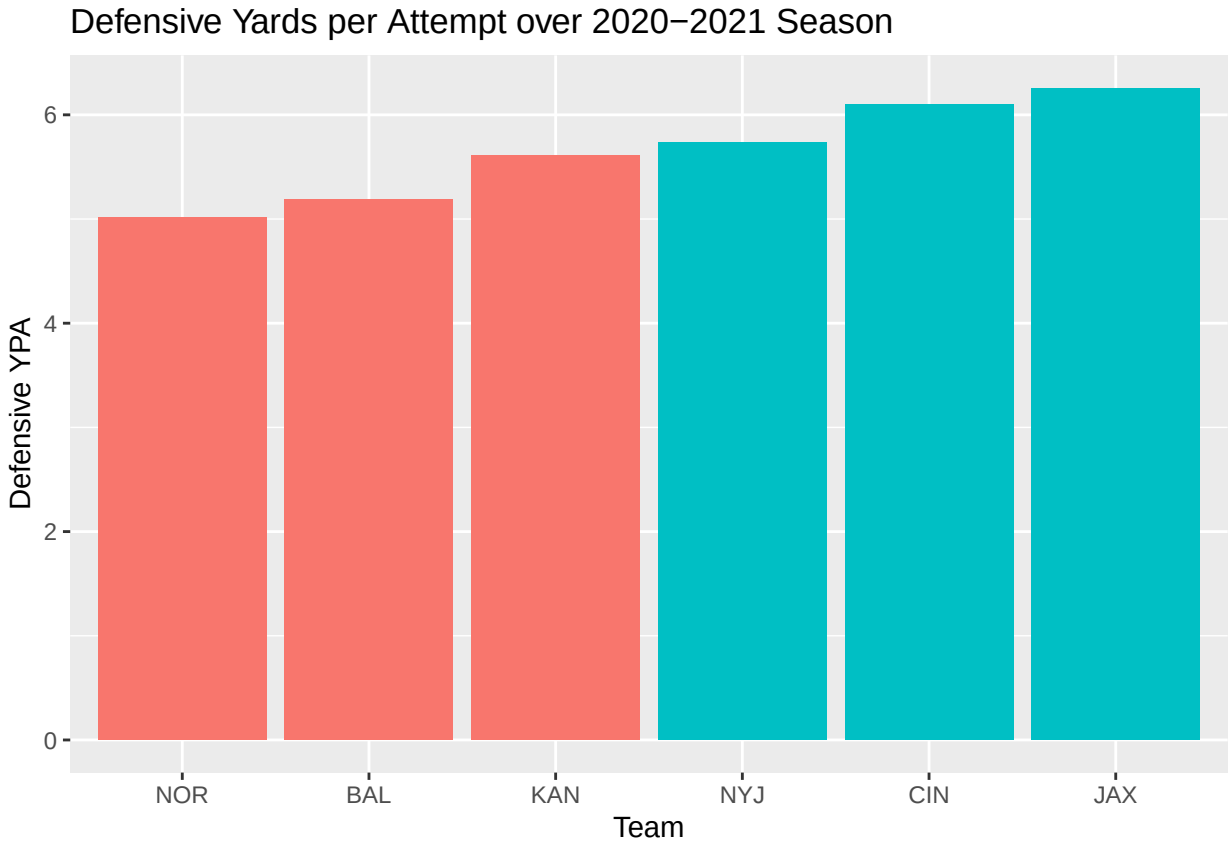
```
ggplot(bestAndWorst, aes(reorder(Tm, `Y/P`), `Y/P`, fill=color)) +
  geom_col() +
```

```
  labs(x="Team",y="Offensive YPA") +
  theme(legend.position = 'none') +
  ggtitle("Offensive Yards per Attempt over 2020-2021 Season")
```



A football team is still a team. There is no one player who can singlehandly win a game or define the odds. The offensive ability of an entire team must be considered when determining a teams chances to win a game. The above plot suggests that oddsmakers would agree.

```
ggplot(bestAndWorst, aes(reorder(Tm, `DY/P`), `DY/P`, fill=color)) +
  geom_col() +
  labs(x="Team",y="Defensive YPA") +
  theme(legend.position = 'none') +
  ggtitle("Defensive Yards per Attempt over 2020-2021 Season")
```



Similarly, a team's defensive ability must be considered. Oddsmakers appear to agree in this case too.

Data Analysis

There are two approaches I want to take when it comes to the linear analysis of this data. First, I want to use what I believe to be important in determining odds as the independent variables. These variables come from my own knowledge as a football fan and the relationships recognized during the previous section. I also want to use AIC to find the "best" model under the AIC criteria.

To start with my first approach, I chose to look at week 3 data. Week 3 was a good week in the NFL because each team played in that week and in the previous week. There are many weeks where some teams did not play due to covid issues or bye weeks.

```r
# Get rid of unneccesary columns
week3odds <- subset(odds, odds$week == 3,
                    select = c('Team','implied','VH'))
# Merged odds and stats
week3merged <- merge(week3odds, week2, by.x='Team', by.y='Tm')
# Create dummys for home team and a win in previous week
week3merged$homeaway <- ifelse(week3merged$VH == 'H', 1, 0)
week3merged$prevWin <- c(1,0,1,1,0,1,0,1,1,0,
                         0,1,0,1,0,1,0,1,1,0,
                         0,0,0,0,0,0,1,1,1,1,
                         1,0)
```

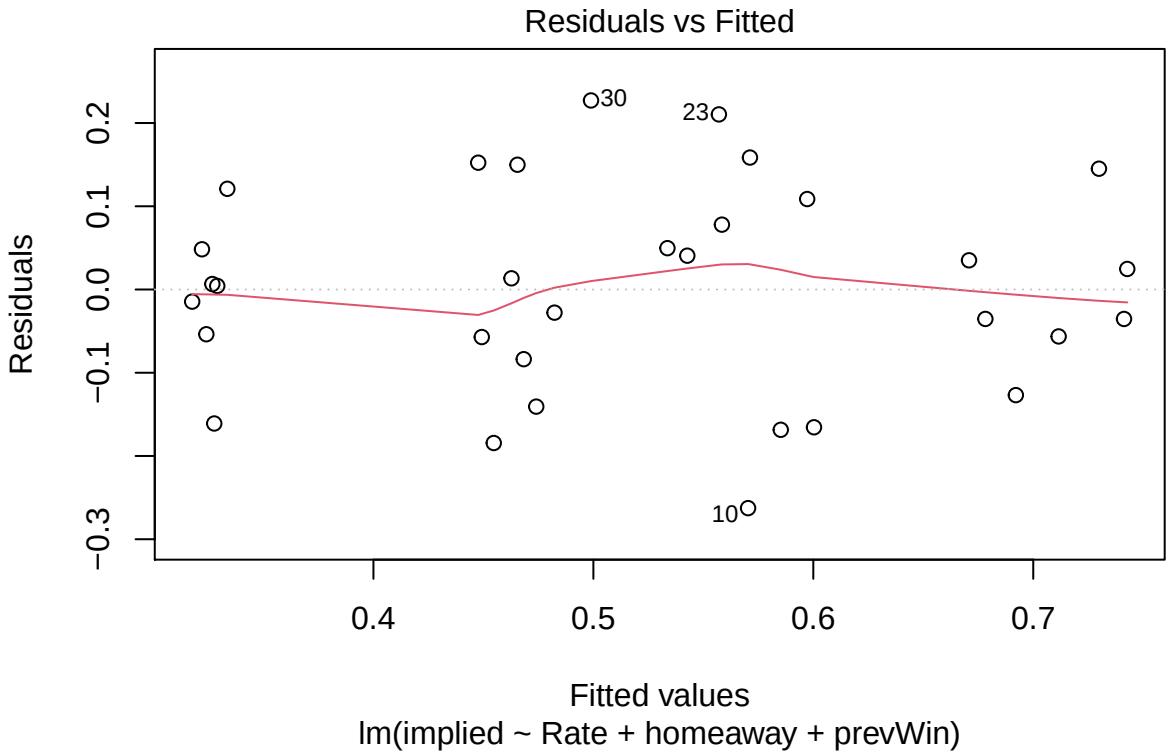
The first model I decided on was an MLR that used Quarterback Rating and two dummy variables. The dummys were encoded as 1 if the team was the home team for the first dummy and as 1 if the team won their last game for the second dummy.

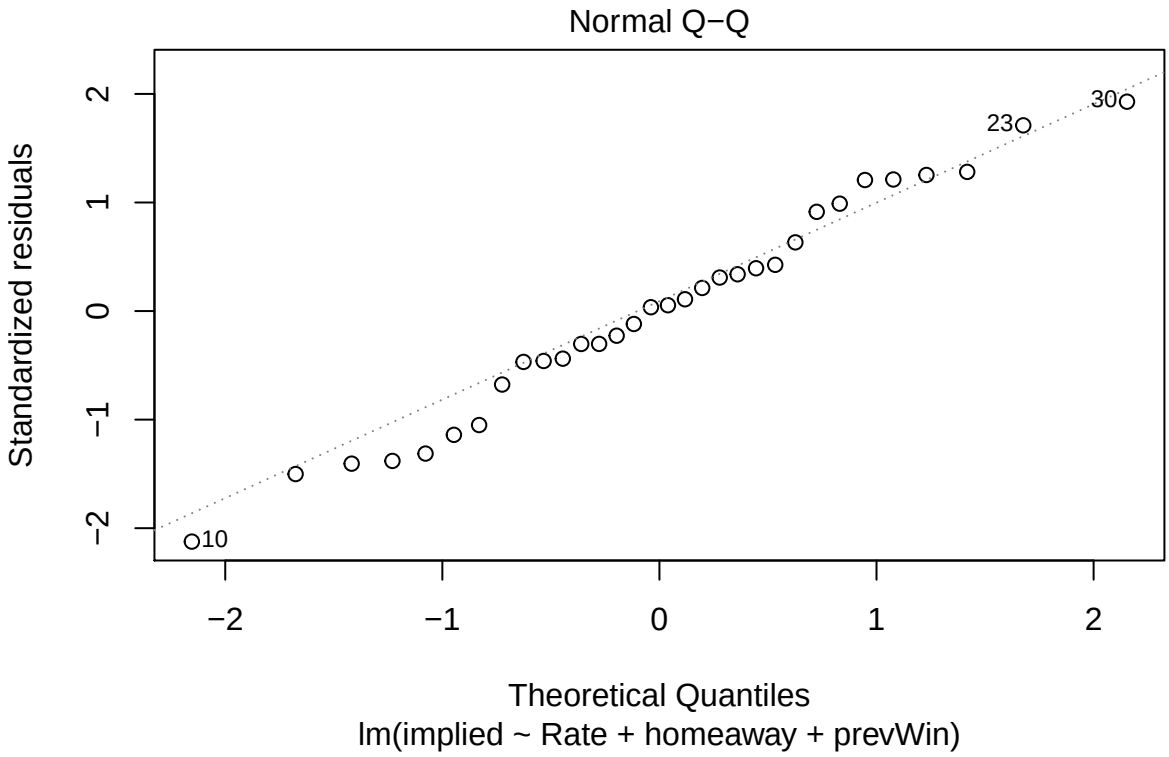
```r
week3mod1 <- lm(implied~Rate+homeaway+prevWin,data=week3merged)
summary(week3mod1)
```

```
##
## Call:
## lm(formula = implied ~ Rate + homeaway + prevWin, data = week3merged)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.262642 -0.063753  0.005461  0.085607  0.227101
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  0.429937   0.130401   3.297  0.00266 **
## Rate        -0.001239   0.001471  -0.842  0.40700
## homeaway     0.247046   0.046435   5.320 1.16e-05 ***
## prevWin      0.167214   0.055559   3.010  0.00548 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1294 on 28 degrees of freedom
## Multiple R-squared:   0.55,  Adjusted R-squared:  0.5018
## F-statistic: 11.41 on 3 and 28 DF,  p-value: 4.607e-05
```

```
plot(week3mod1, which=c(1,2))
```





Surprisingly, in this model, QBR was not statistically significant. However, the two dummys both were significant at the 0.01 level. The β estimate for home/away was 0.247 meaning that holding other variables constant, you would expect an increase in implied odds by 0.247 when a team is at home vs. when they are on the road. The β estimate for a win in the previos week was 0.167, meaning that holding other variables constant, you would expect an increase in implied odds by 0.167 when a team has won its previous game vs. when it did not. This effect is less than the home/away effect but still significant.

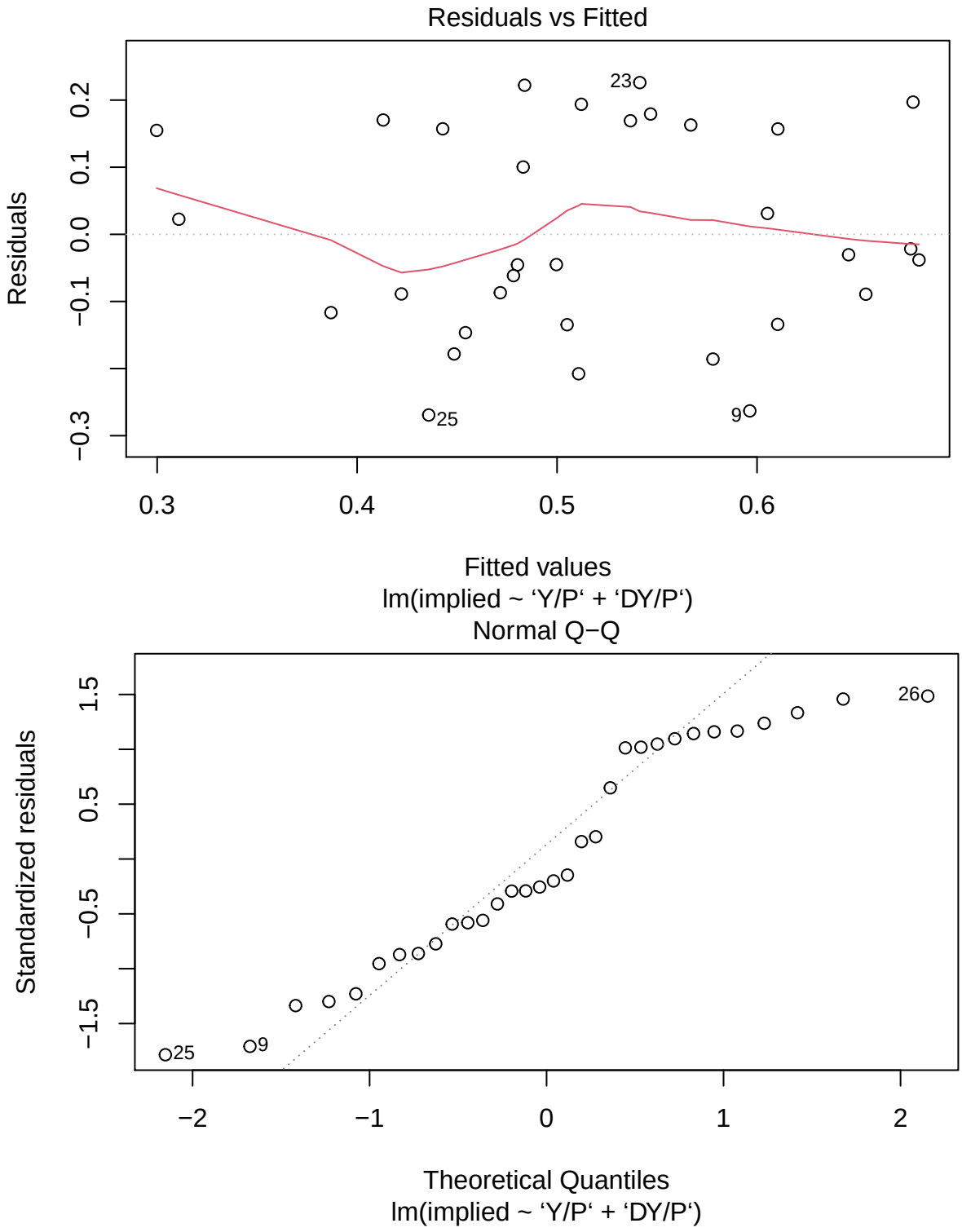
The diagnostic plots for this model look good. The residuals vs. fitted line is close to horizontal and the points follow the Normal Q-Q plot closely, indicating that a linear relationship is present.

The second model I tried was one that took into account a teams offensive ability (Yards per play) and defensive ability (Defensive yards per play). This made sense because they were on the same scale.

```r
week3mod2 <- lm(implied~`Y/P`+`DY/P`,data=week3merged)
summary(week3mod2)
```

```
##
## Call:
## lm(formula = implied ~ `Y/P` + `DY/P`, data = week3merged)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.26914 -0.12103 -0.03428  0.15859  0.22605
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.57807    0.30529   1.894  0.06831 .
## `Y/P`        0.10138    0.04092   2.478  0.01930 *
## `DY/P`      -0.11199    0.04059  -2.759  0.00994 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1591 on 29 degrees of freedom
## Multiple R-squared:  0.2957, Adjusted R-squared:  0.2471
## F-statistic: 6.087 on 2 and 29 DF,  p-value: 0.006204
```

```r
plot(week3mod2, which=c(1,2))
```



In this model, both variables were statistically significant at the 0.05 level. The β estimate for OY/P indicates that for each extra yard an offense gains per play, you would expect their implied odds to increase by 0.101 when not considering their defense. The β estimate for

DY/P indicates that for each extra yard an defense gives up per play, you would expect their implied odds to decrease by 0.112 when not considering their offense. Both of these varaibles appear to lead to big swings in implied odds.

Though they are not as strong as the previous model, the diagnostic plots for this model look good. The residuals vs. fitted line is close to horizontal and the points follow the Normal Q-Q plot closely, indicating that a linear relationship is present.

We can use this model to try to predict what the implied odds would look like for a superteam (one that gains 7 yards per play and gives up only 4) and for a bad team (one with the opposite stats).

```
# Prediction interval for a superteam
predict(week3mod2, newdata=data.frame('Y/P' = 7,'DY/P'= 4, check.names = F),
        interval = 'prediction')
```

```
##         fit       lwr      upr
## 1 0.8397389 0.4591011 1.220377
```

```
# Prediction interval for a bad team
predict(week3mod2, newdata=data.frame('Y/P' = 4,'DY/P'= 7, check.names = F),
        interval = 'prediction')
```

```
##         fit        lwr       upr
## 1 0.1996257 -0.1809616 0.5802129
```

The superteam's prediction interval is centered around 0.8397 and goes over 1 in its upper bound, suggesting that a team with these stats would be a given insanely high odds. The bad team's prediction interval is centered around 0.1996 and goes below 1 in its lower bound, suggesting that a team this bad would be given an abysmal chance to win.

Now, let's see what AIC would have selected:

```r
# Remove some colinear (Ex: Pts For, Pts Against vs. Pt Diff)
# and non-useful columns (Ex: Week, Team)
week3useful <- subset(week3merged, select = -c(1,3,4,6,7,10,15,16))

# Step forward
noVars <- lm(implied~1, data=week3useful)
allVars <- lm(implied~., data=week3useful)
step(noVars, scope=list(lower = noVars, upper=allVars), direction = "forward")
```

```
## Start:  AIC=-107.57
## implied ~ 1
##
##             Df Sum of Sq     RSS     AIC
## + homeaway   1   0.40469 0.63794 -121.29
## + `DY/P`     1   0.15285 0.88977 -110.64
## + `Y/P`      1   0.11555 0.92707 -109.33
## + prevWin    1   0.09826 0.94436 -108.73
## + `W-L%`     1   0.09467 0.94795 -108.61
## + PD         1   0.08700 0.95562 -108.36
## <none>                   1.04262 -107.57
## + TOM        1   0.02619 1.01643 -106.38
## + Rate       1   0.02164 1.02098 -106.24
## + PTD        1   0.01612 1.02650 -106.07
## + RTD        1   0.00347 1.03915 -105.67
##
## Step:  AIC=-121.29
## implied ~ homeaway
```

```
##
##           Df Sum of Sq     RSS     AIC
## + prevWin  1  0.156884 0.48105 -128.32
## + `W-L%`   1  0.094668 0.54327 -124.43
## + PD       1  0.066244 0.57169 -122.80
## + `Y/P`    1  0.052293 0.58564 -122.03
## + `DY/P`   1  0.045389 0.59255 -121.65
## <none>                 0.63794 -121.29
## + TOM      1  0.020170 0.61776 -120.32
## + Rate     1  0.016976 0.62096 -120.15
## + RTD      1  0.013800 0.62414 -119.99
## + PTD      1  0.007160 0.63077 -119.65
##
## Step:  AIC=-128.32
## implied ~ homeaway + prevWin
##
##           Df Sum of Sq     RSS     AIC
## <none>                 0.48105 -128.32
## + PTD      1 0.0146116 0.46644 -127.31
## + Rate     1 0.0118754 0.46918 -127.12
## + `DY/P`   1 0.0110221 0.47003 -127.06
## + RTD      1 0.0073498 0.47370 -126.81
## + `W-L%`   1 0.0018051 0.47925 -126.44
## + PD       1 0.0008373 0.48021 -126.38
## + `Y/P`    1 0.0005358 0.48052 -126.36
## + TOM      1 0.0002725 0.48078 -126.34
##
```

```
## Call:
## lm(formula = implied ~ homeaway + prevWin, data = week3useful)
##
## Coefficients:
## (Intercept)     homeaway      prevWin
##      0.3259       0.2426       0.1411
```

Stepping forward with AIC shows that AIC is minimized with the home/away dummy and the previous win dummy. The coefficients for this model show that given when holding the previous week win variable constant, being the home team, you would exepct to upgrade a team's implied odds by 0.2426. They also show that holding the home/away variable constant, when winning in the previous week, you would expect a team's implied odds to increase by 0.1411. This is very similar to the first model I proposed, which has the same significant variables as AIC's model.

```
confint(lm(implied~homeaway+prevWin, data=week3useful))
```

```
##                  2.5 %    97.5 %
## (Intercept) 0.24145374 0.4102678
## homeaway    0.14868885 0.3364230
## prevWin     0.04727751 0.2350117
```

The confidence interval for this model shows that it would be reasonable for the effect of being the home team on implied odds to be anywhere between 0.1487 and 0.3364. This upper bound is very impactful, suggesting that oddsmakers definitely believe in home field advantage. The confidence interval for the previous win variable shows that it would be reasonable for the effect of winning your last game to be anywhere between 0.0473 and 0.2350. If the lower bound were true, this effect would be marginal.

Summary and Discussion

There are many takeaways from thislinear regression analysis. Firstly, one of my most confident predictions was that a team's quarterback rating was going to make a significant linear impact on a team's implied odds. The regression analysis said otherwise and completely rejected my hypothesis. I was not all wrong though. I had a feeling that recent results and home field advantage played a measurable role in the odds determined by oddsmakers. This was proven to be true by my predicted model and the model produced by stepping forward with AIC. The optimal model produced by AIC did vary from my second predicted model which showed statistical significance in offensive yards per play and defensive yards per play. This suggests that those two stats may not be as impactful as I thought but does not confirm it. One of the main limitations of this analysis was inability to consider who a team was playing as a part of the model. Due to limitations in the data that I could find, it was hard to match up each team's opponent. Given more time and better data, I think the stats of a team's opponent would have some sort of significance in that team's implied odds. Another limitation is the lack of data that I had to do my analysis compared to what oddsmakers most likely use. I believe that oddsmakers have supercomputers that run much more sofisticated models than I could possibly create with my data. Future research should compile more data than I was able to in order to gain more insight on what oddsmakers are using to make their NFL picks.

References

- https://www.sportsbookreviewsonline.com/scoresoddsarchives/nfl/nfloddsarchives.htm for odds data
- https://www.pro-football-reference.com for weekly stats

Data and Code

All data and code can be found at https://github.com/RossBechtel/STAT410FinalProject