NFL Home Team Advantage Analysis

2023-12-17

Introduction

The National Football League (NFL) is one of the biggest sport leagues in our country. It has a long history, dating back to even the 1930s, and it is a sport that many people can enjoy watching even without knowing too much about it. It is normal for people to have a favorite team, and whether that team is relatively good or not is usually a factor for new watchers. In this project, I want to look into the phenomenon known as home field advantage. I'll be examining how much statistical truth there is the the claim that teams win more games at home than they do away, relative to their overall performance. I want to discover if there is a relationship between the amount of games won at home vs. away and how teams perform in the different scenarios.

By looking into factors like win/loss ratio and the average margin of victory (MOV), I want to see if the comfort and familiarity of the home field, along with (usually) a much larger proportion of fans rooting for teams at home, makes a difference in the outcome of games.

The overall aim of this project is to shed light on the competitiveness of the NFL, and give a deeper understanding of the factors that go into how teams play and the outcome of games.

Data Description

I gathered the data from a website called "teamrankings.com". The data includes every team's amount of wins, losses, ties, win percentages, and average margins of victory for all games. I have data from every home and away game dating back to the 2003 season, as this was as far back as the database went. I think 20 seasons of data is a good amount because it goes back far enough to have enough data, and also captures relatively how good teams are, which is important when accounting for win rates home vs away. I also have a variable in the data called ATS +/- , but I am not using this variable because it doesn't relate to the teams' wins, it is just a measure for betting. While it does relatively show something about how much teams win/lose by, the MOV variable already captures this so it is not necessary. Here is a quick look into the data, where the first set is from away games, and the second is from home games:

```
##           Team Win.Loss  Win..  MOV ATS.... WinRate Wins Losses
## 1  New England 110-62-0 64.00%  5.7     2.8    64.0  110     62
## 2   Pittsburgh  97-76-1 56.10%  0.6     0.0    56.1   97     76
## 3 Philadelphia  96-77-1 55.50%  1.0     0.7    55.5   96     77
## 4 Indianapolis  96-80-1 54.60% -0.3     0.6    54.6   96     80
## 5  New Orleans  92-82-0 52.90%  1.3     1.4    52.9   92     82
## 6    Green Bay  92-86-0 51.70%  0.4     0.6    51.7   92     86
```

```
##          Team Win.Loss.Record  Win.. MOV ATS.... WinRate Wins Losses
## 1   Green Bay         89-27-2 76.70% 8.2     2.1    76.7   89     27
## 2 New England         95-30-0 76.00% 9.8     3.0    76.0   95     30
## 3   Baltimore         81-34-0 70.40% 7.0     2.0    70.4   81     34
## 4     Seattle         82-36-0 69.50% 6.0     1.5    69.5   82     36
## 5  Pittsburgh         80-36-1 69.00% 5.9     2.1    69.0   80     36
## 6 Kansas City         83-40-0 67.50% 5.1     0.1    67.5   83     40
```

To get this data, I had to copy it from the website I first put it into an excel sheet. I couldn't simply download it as a csv or spreadsheet. Once in the spreadsheet, all I had to do was save it as a csv file, and import it into R using the read.csv() command. Once in a data frame, I extracted the wins, losses, and win percentage as numbers because they were represented with strings, and I couldn't work with them like that. The 3 last columns - WinRate, Wins, and Losses are from my cleanup to make the data easier to work with. Since the MOV column was already a double, I didn't have to do anything with it. I don't see any potential issues with this data, the only thing I can say may be a problem is the lack of depth and context around the data. If there was a way I could get more information about the data, or maybe more measures of how the games turned out, I think that would benefit me and make my analysis better.

Methods

I used various methods from class to test my data. I went about this approach by considering what the important things are to show that prove/disprove a statistical significance of a home field advantage. My initial approach was to fit a linear regression of the away wins/losses on the home wins/losses to see if they could be reliably predicted. I chose this approach to start because it gives a good basis to show how significant the differences are and how much of a pattern is followed. It also gave me a good idea of how I wanted to do the further analyses, as I knew what questions were left to be answered.

Next, I used a paired t-test as we learned to test if there was a difference between 2 variables. In my case, I used this method to see if there was a significant difference between the margin of victory (MOV) between the home and away statistics. This test goes through all 32 teams and subtracts the away MOV from the home MOV to see if there is a true difference. I used another paired t test on the win rates between home and away. The reason I used it again is because I feel it's the strongest way to see if there is a true difference between the two, which is the overall goal of this project.

Nearing the end, I did another linear regression, fitting the away win rate on the home win rate, to see how well it could be predicted. Lastly, I plotted the difference between the amount of wins away and the amount of wins at home, and the same for losses. This was to hammer home the point I'm trying to make in this project - plotting their differences can show a clear trend, which could be backed up by the statistical tests I have mentioned before.

Results

Here comes the good part. Below I will show results of my tests and what they mean for my hypothesis.

```
##
## Call:
## lm(formula = home_stats$Wins ~ away_stats$Wins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2884 -2.6108 -0.6588  1.1737 13.3008
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.96856    3.45595   1.148     0.26
## away_stats$Wins  0.80135    0.04548  17.622   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.978 on 30 degrees of freedom
## Multiple R-squared:  0.9119, Adjusted R-squared:  0.909
## F-statistic: 310.5 on 1 and 30 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = home_stats$Losses ~ away_stats$Losses)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5174 -1.2637  0.0412  1.5460  4.6777
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -19.60951    3.00031  -6.536 3.15e-07 ***
## away_stats$Losses   0.71220    0.03059  23.279  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.349 on 30 degrees of freedom
## Multiple R-squared:  0.9475, Adjusted R-squared:  0.9458
## F-statistic: 541.9 on 1 and 30 DF,  p-value: < 2.2e-16
```

Above, I have two summaries printed about predicting the amount of wins and losses at home based on the number from away games. Looking at the summaries, we see the slopes for both are extremely significant, which leads to the conclusion that they are different from 0. Knowing this, I can conclude that both can, to a degree, predict the other, which shows a pattern and helps get the ball rolling on the idea that there is a true home field advantage. We see that on average, as the number of wins away increases by 1, the increase by .8. The number of losses at home increases by .71, though. While this may not be extremely convincing, the later tests will show this to be significant.

```
##
##  Paired t-test
##
## data:  home_stats$MOV and away_stats$MOV
## t = 14.574, df = 31, p-value = 2.035e-15
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  3.754689 4.976561
## sample estimates:
## mean difference
##        4.365625
```

Above I have the summary of a paired t test between the MOV home and away. We see here that there is a p-value of 2.03e-15, meaning this is extremely significant at the 0.05 level. As shown in the summary, we reject the null hypothesis that the true mean difference is equal to 0, and come to the conclusion that there is a true difference between these values. We see that the mean difference from our data is 4.366. This helps build the case even more that there is a home field advantage. Since we have 20 years of data and the p value is extremely low, it is very unlikely that this result is just due to chance and the teams getting lucky schedules. This shows us that teams actually do perform better at home on average, and by a pretty large margin.

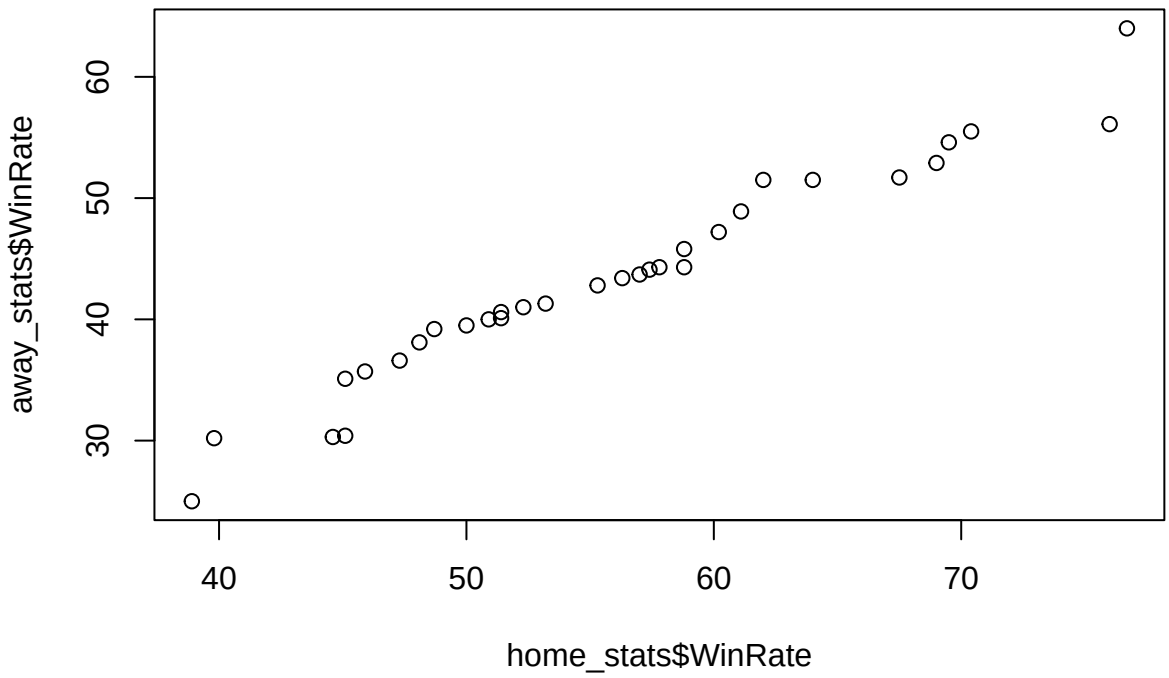
```
##
##  Paired t-test
##
## data:  home_stats$WinRate and away_stats$WinRate
## t = 31.369, df = 31, p-value < 2.2e-16
```

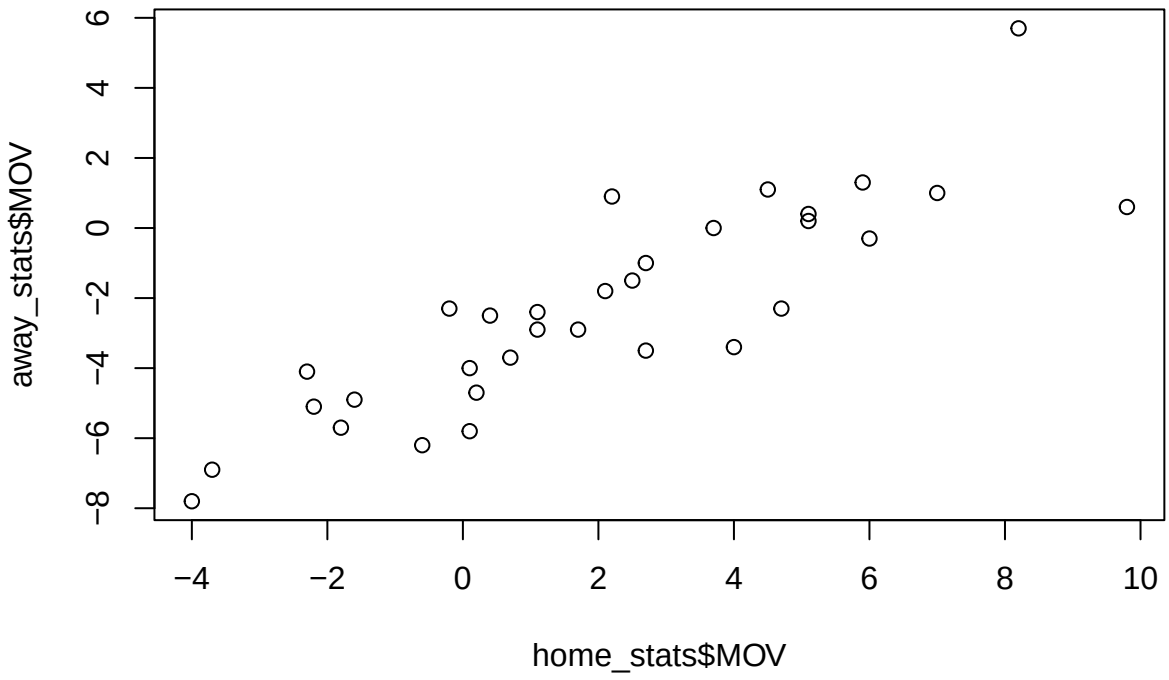
```
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  11.83631 13.48244
## sample estimates:
## mean difference
##        12.65937
```

Here is another paired t test between the home and away win rates. This is where a conclusion can confidently be made that there is a home field advantage. We see a p-value below 2.2e-16, so low that it is almost guaranteed this is not just chance. We reject the hypothesis that the true mean difference is equal to 0, and conclude there is a difference. In fact, according to the test, with 95% confidence, this difference is between 11.8 and 13.5 percent higher win rates at home. This is a massive difference, and undeniably shows that there is a home field advantage.

```
##
## Call:
## lm(formula = home_stats$WinRate ~ away_stats$WinRate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1030 -1.5507 -0.3193  0.8629  5.7681
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         7.68112    1.90858   4.025 0.000357 ***
## away_stats$WinRate  1.11499    0.04325  25.779  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.088 on 30 degrees of freedom
## Multiple R-squared:  0.9568, Adjusted R-squared:  0.9554
## F-statistic: 664.5 on 1 and 30 DF,  p-value: < 2.2e-16
```

This is another linear model regressing away win rates on home win rates. With this, we see the slope is very significant, at the .001 level, and that with every increase by 1 of the away win rate, we predict the home win rate goes up by 1.11, which is again, an important difference. It is also important to note the R^2 of this and the previous regression models, as they are all strong at levels above .9. This shows that our models have fit very well, strengthening the idea of a home field advantage.


away_stats$WinRate
30
40
50
60
40
50
60
70
home_stats$WinRate



Lastly, above there are two graphs plotting the difference between home and away wins/losses. These graphs are just the cherry on top to visualize a clear relationship between the home and away statistics. Looking at these graphs, of course as teams are better and win more at home or win by more at home, they will do the same away. But that is not the point of these graphs. If you pay attention, you'll see that for any given point on either graph, the win rate or MOV at home is larger than that of away. That is the point in these graphs. There is not a single outlier - every single point follows that pattern.

Conclusion

My analysis in this project has led to the conclusion that there is in fact a home field advantage present in the NFL. Looking at all of the tests, every test I did had significant results and there is proof beyond a reasonable doubt based off of these statistics. While I have not answered exactly why it is, I have shown that it undoubtedly exists. This trend shows the power of what I imagine to be familiarity and fan support.

In the future, studies could go deeper into external variables, such as weather conditions, crowd support, the comfort of being at home, travel fatigue, etc. It would be interesting to see what it actually is contributing to this large difference. I would like to know the psychological differences of players at home vs away. Understanding more of the causes can also help case of the existence of a home field advantage, and perhaps the NFL can take measures to curb this advantage if possible or if they see fit. To conclude this, it is also possible that examining these external variables could lead to a discovery opposing what I have found here.

Overall, this project discovered a clear home field advantage in the NFL, seeing that teams win much more often and by larger margins at home. This is important knowledge in predicting the outcome of games, as if two teams are relatively evenly matches, it is more likely that the team at home will win. However, it is important to understand the limitations. This conclusion should not be used to predict that the 0-15 Browns will beat the 12-3 Patriots (just an example) just because they are at home. There are many other factors

that go into the outcome of games. With that said, it can help to understand the outcomes of games better and can have an application in predicting winners of games.