Dean&Francis

ISSN 2959-6157

# The Effect of NBA Attendance on Teams' Performance

**Chenzhao Wang**

Tandon/CAS, New York University, New York, 10003, US
Corresponding author email: cw3544@nyu.edu

**Abstract:**

This paper analyzed how attendance in the NBA during the COVID-19 pandemic, specifically, in the 2020-2021 seasons, affected team performance. To answer this question, we used team data of all 30 teams from 2015 to 2022, and attendance data for the 2020-2021 season; we separated team attendance into three levels: Full (or almost full), Partial, and Empty. We used logistic models and the Bradley-Terry Model to analyze each team's strength coefficient (compared to one baseline team) using the Win/Loss data as the determining factor. Our result showed a slight advantage for home teams when audiences were present. However, results revealed that home court advantage became less evident when there was partial or no audiences.

**Keywords:** Attendance, Regression Model, Strength Coefficient, Covid, Bradley-Terry Model

## 1. Introduction

Starting from early 2020, many sports events around the world were forced to shut down because of the COVID-19 pandemic. On July 31, 2020, the National Basketball Association (NBA), a professional basketball league in North America comprised of 30 teams (29 from the U.S. and 1 from Canada), moved to closed arenas in Orlando, Florida called the "Bubbles." [1]

During that period and most of the 2020-2021 season, no audiences or very limited audiences were allowed in the arenas to watch the games in person. However, it was important to acknowledge the fact that the audience always played a critical role in all sports since they would cast both positive and negative effects on both teams.

Previous studies have analyzed the effect of audiences on NBA star players, however, not on the effect of the whole team as a perspective by Cory Metcalfe in 2013[2]; and by Humphreys and Johnson [3]. A similar study, using two stage least-squares, discovered that there existed a linear, negative correlation between attendance and away teams' free throw percentage, by La in 2014.[4]

This research examined and answered the following question: What effects do audience attendance have on the performance of NBA teams?

Our group gathered raw CSV files from Basketball Reference, a branch of Sport Reference LLC [5], which provided raw CSV files of all NBA teams and player statistics since the creation of the league.

This report was organized into sections, with section 2 introducing the data set, data cleaning, and data merging, section 3 presenting data exploration and analysis with graphs and logistic models, and section 4 concluding the report.

## 2. The data

• Detailed in-game statistics of all 30 teams from the 2015-2016 to 2021-2022 seasons. (Field Goal Percentage, Free Throw Made, etc.)
• Attendance numbers during the 2020-2021 COVID-19 season.

## 3. File handling

### 3.1 *Read in files*

Read in all NBA team stats files from 2015-2021 and attendance data for the 2020-2021 pandemic season. We gathered attendance data from only the 2020-2021 season based on the assumption that the normal seasons were played with mostly full audience attendance.

### 3.2 *Modify, clean up, and data merging*

We had two groups of files: Team Data and Attendance Data. We wanted to extract the attendance numbers from Attendance Data and merge them into Team Data, which will be a long final data frame.

We modified and beautified column names for all files. (Eg. Separated home team stats and opponent stats; Turned special character in 3P% to "ThreePP" (Three-pointer percentage))

We assumed that the attendance numbers for games in the non-COVID season were almost full. By experience, we set those numbers to 25,000, which is approximately the maximum capacity of NBA stadiums.

*Important*: Thanks to Professor Emerson, we found out that the column named "Home" ("X" in raw data) refer-

eed to whether the team was at home or visiting. ("@" = Visiting, "" = Home) Thus, we were able to get rid of the duplicate games in the data set since a game would appear two times - one from the Home team file and one from the Visitor team file.

Next, we merged the two different files into one final data set containing game stats from 2015-2022 and attendance numbers for the 2020-2021 season.

From the table (excerpt) below, we discovered that there were 7 games in the 2020-2021 season for which the audience number was not recorded (represented by empty strings). By checking the video recordings of those games, we found out there were actually 0 audiences in the arenas.

**Table 1 Records of NBA games attendance in the 2020-2021 season**

| | 0 | 1,000 | 1,008 | 1,180 |
|---|---|---|---|---|
| 7 | 570 | 3 | 1 | 1 |

Moreover, some games in the 2019-2020 season were played in closed-to-public arenas (Bubbles) in Orlando, Florida at the Walt Disney World from 2020/07/31 to 2020/08/14.

Therefore, we set those attendance numbers to 0 and the home variable to "N" for the Neutral site.

In the end: the group modified home team names to match the official abbreviations (Eg. "Boston Celtics" to "BOS") to facilitate future data exploration. Additionally, the names of Opponent teams in the original data file were also in abbreviations, so this gave us a nice consistency.

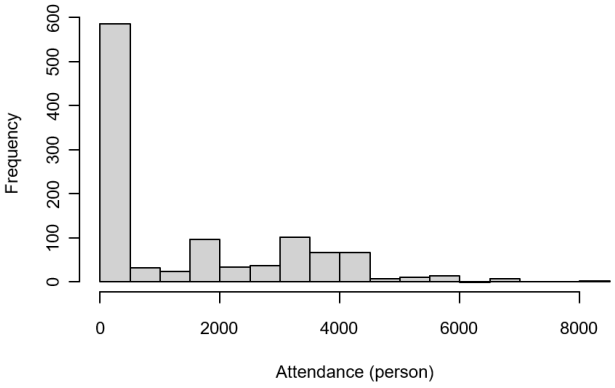


**Figure 1 Histogram of attendance data in the 2020−2021 COVID season**

## 4. Data exploration and analysis

### 4.1 *Basic Explorations*

First of all, we looked at a histogram of attendance in the 2021-2022 COVID season. As expected, there were almost 600/1080 games where there were no audiences at all, so this year would be a good fit for our research because there was a approximately even spread between no audiences and limited audiences.

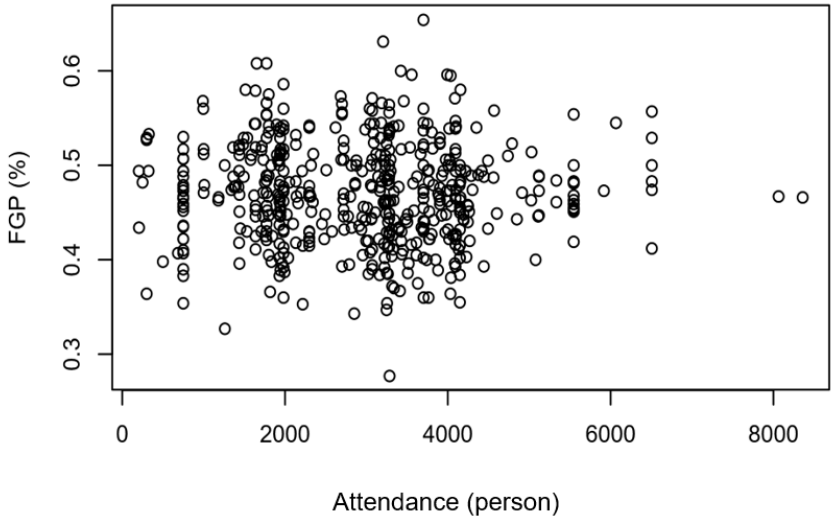


**Figure 2 Scatter plot of NBA game Attendance to Field Goal Percentage in the 2020−2021 COVID season**

Next, to see if attendance had an impact on team performance, we chose FGP (Field Goal Percentage) as the dependent variable and made a plot. There appeared no significant trend in the scatter plot. Moreover, a linear model of the same variables demonstrated that the slope of this scatter plot was nearly 0, meaning that there were virtually no significant correlations between attendance and FGP.

Furthermore, by intuition, we believed that home-court advantage played a significant role in close games because the audience can influence the performance of both teams. For this study, we defined a close game as one with a ten-point difference in the last three minutes. However, the home court advantage was expected to be minimal during the pandemic as there was no limited audience.

| Full-Attendance | Partial or Empty | Covid-Period |
|---|---|---|
| 0.5206612 | 0.5265306 | 0.5042735 |
| 0.5442677 | 0.5467626 | 0.5384278 |

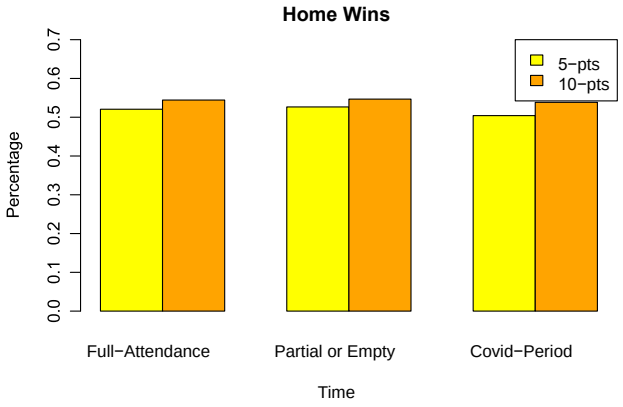


**Table 2 Percentage of home wins varied by different attendance levels in close games**

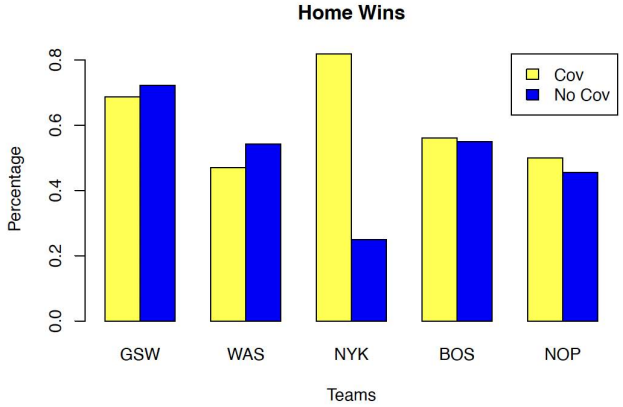


**Figure 3 Histogram of NBA home game winning percentage in close games regarding Attendance in three criteria**

However, the percentage varied slightly throughout different times both by the table and the plot. To be more specific, we decided to narrow down the time scope and teams for a more thorough analysis. To minimize the impact of team changes, we analyzed the home court performance of five teams in the 18-19 season, the season prior to the pandemic, by two high-performing teams, two low-performing teams, and one mediocre team. We compared their pre-COVID and COVID home court performance, and the result is revealed by the following plot.

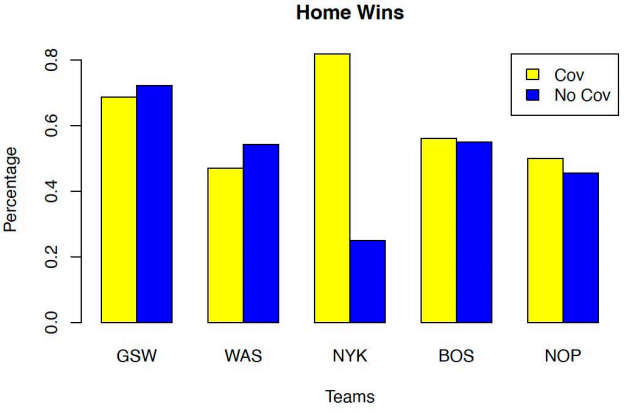


**Figure 4 Histogram of NBA home game winning percentage in close games for selected five teams during and before 2020-2021 COVID season**

According to the histogram, the result misaligned with our intuitions that the difference was insignificant. It showed that two teams performed better pre-COVID, and the other two changed slightly. NYK 's roster underwent vast changes through drafts and trades, so their performance was improved.

One of the most important variables that would affect game results is a team's strength level, which could be computed with Field Goal Percentage, Three-Pointer Percentage, etc. However, we thought the previous game results (win or loss) would be a better indicator of team strength since it'd be more direct than other game factors.

Now that we established a determining factor of a team's strength level, we used the Bradley-Terry Model [6] to help us estimate team strength coefficients over the seven

chosen seasons.

### 4.2 *A preliminary model with past game histories*

Before computing the strength coefficients for 30 teams with logistic model, we initialized a matrix filled with 1 for Home and -1 for Visiting for all teams from 2015-2022. We made sure that every row would sum up to zero, assuring us that each game was played between two different teams and that the matrix was safe to use.

When analyzing the data, we excluded WAS (the Washington Wizards) and used it as a *baseline team* for other teams to compare with. The WAS roster did not change that much over the years as the others, so its strength coefficients over the selected seasons would remain relatively consistent [7]. Therefore, to contrast the strength coefficients for different teams over seasons, we used the remaining 29 teams for modeling. To better understand and quantify the strength coefficients, we first used data from the 2015-2016 season and made a generalized linear model as an example. This helped us check the validity of the logistic model by comparing the strength coefficients to the actual team rank and performance that year. For example, the Golden State Warriors did a fabulous job and ranked number one in the league and the Western Conference in the 2015-2016 season, so we should be expecting a relatively high strength coefficient.

**Table 3 Estimation of strength coefficients by using the game results from 2015-2016 season**

| | | |
|---|---|---|
| (Intercept) 0.46258 | | 0.06688 6.916 4.64e-12 *** |
| ATL | 0.39782 | 0.33377 1.192 0.233291 |
| BOS | 0.35134 | 0.33436 1.051 0.293354 |
| BRK | -1.21502 | 0.35891 -3.385 0.000711 *** |
| CHI | 0.05494 | 0.33436 0.164 0.869495 |
| CHO | 0.35297 | 0.33434 1.056 0.291094 |
| CLE | 0.88013 | 0.34550 2.547 0.010854 * |
| DAL | 0.09130 | 0.33944 0.269 0.787946 |
| DEN | -0.44273 | 0.34254 -1.293 0.196181 |
| DET | 0.14621 | 0.33222 0.440 0.659862 |
| GSW | 2.30041 | 0.43973 5.231 1.68e-07 *** |
| HOU | 0.01596 | 0.33852 0.047 0.962392 |
| IND | 0.22368 | 0.33416 0.669 0.503264 |
| LAC | 0.68671 | 0.34767 1.975 0.048251 * |
| LAL | -1.47835 | 0.37603 -3.931 8.44e-05 *** |
| MEM | 0.05195 | 0.33977 0.153 0.878490 |
| MIA | 0.38553 | 0.33355 1.156 0.247744 |
| MIL | -0.43111 | 0.33464 -1.288 0.197643 |
| MIN | -0.68785 | 0.34485 -1.995 0.046082 * |
| NOP | -0.63172 | 0.34496 -1.831 0.067060 . |
| NYK | -0.52281 | 0.33929 -1.541 0.123337 |
| OKC | 0.81812 | 0.35018 2.336 0.019477 * |
| ORL | -0.35460 | 0.33390 -1.062 0.288243 |
| PHI | -2.15497 | 0.42045 -5.125 2.97e-07 *** |
| PHO | -1.07643 | 0.35711 -3.014 0.002576 ** |
| POR | 0.18383 | 0.33957 0.541 0.588268 |
| SAC | -0.45403 | 0.34204 -1.327 0.184368 |
| SAS | 1.69088 | 0.38929 4.343 1.40e-05 *** |

| | | |
|---|---|---|
| TOR | 0.82713 | 0.34313 2.411 0.015929 * |
| UTA | -0.05294 | 0.33950 -0.156 0.876094 |

The summary showed 29 teams and their strength coefficients. We could clearly see that the GSW had a significantly strong strength coefficient of about 2.3, which matched the team's actual performance that year, proving the effectiveness of the model. (WAS was set to be the baseline team with a strength coefficient of
0).

Seeing the model worked for the 15-16 season, we proceeded to create a new matrix to store the strength coefficients for all 7 seasons.

The new matrix also stored the "Home Intercept," the intercept of the generalized linear model, which was the predicted value of the dependent variable when all the independent variables were 0. In the case of our research, the "Home Intercept" meant the probability for a Home team to win in a chosen data set when we disregarded the factor of teams and considered only the Home/Visitor factor. Thus, we concluded that a positive intercept would mean that there existed a home court advantage in the data set.

**Table 4 Prediction of percentages of games over seasons**

| Season | Predictions |
|---|---|
| 16-17 | 0.6357724 |
| 17-18 | 0.6138211 |
| 18-19 | 0.6300813 |
| 19-20 | 0.618508 |
| 20-21 | 0.5472222 |
| 21-22 | 0.6081301 |

The above numbers were the percentages of the predicted outcomes of games in a season over the actual outcomes of the games in that season. We observed that in the fifth result, which was the 2020-2021 pandemic season, the percentage was about 0.55, nearly a random guess. Other seasons, either normal or almost normal, all had a solid above-sixty-percent prediction success rate. This meant that there must be some other variable(s) other than solely the past game results that affected the effectiveness of the prediction. At the end of creating the new matrix, we rounded the strength coefficients to 3 decimal places for better representations.

**Table 5 Prediction of strength coefficients for 29 NBA teams over seasons**

| Seasons | 15-16 16-17 17-18 18-19 19-20 20-21 21-22 |
|---|---|
| Home Intercept | 0.463 0.396 0.384 0.453 0.272 0.204 0.207 |
| ATL | 0.398 -0.323 -1.062 -0.143 -0.110 0.412 0.419 |
| BOS | 0.351 0.230 0.637 0.953 1.577 0.116 0.835 |
| BRK | -1.215 -1.593 -0.825 0.573 0.709 0.838 0.479 |
| CHI | 0.055 -0.412 -0.876 -0.564 -0.089 -0.176 0.571 |
| CHO | 0.353 -0.661 -0.396 0.423 0.073 -0.059 0.425 |
| CLE | 0.880 0.137 0.336 -0.773 -0.228 -0.737 0.475 |
| DAL | 0.091 -0.776 -1.034 0.181 1.157 0.504 0.888 |
| DEN | -0.443 -0.414 0.177 1.341 1.453 0.816 0.677 |
| DET | 0.146 -0.607 -0.218 0.519 -0.220 -0.878 -0.661 |
| GSW | 2.300 1.218 0.855 1.495 -0.429 0.329 0.984 |
| HOU | 0.016 0.421 1.326 1.274 1.353 -1.067 -0.879 |
| IND | 0.224 -0.379 0.259 0.886 1.317 0.000 -0.535 |

| | |
|---|---|
| LAC | 0.687 0.195 -0.028 0.984 1.713 0.814 0.392 |
| LAL | -1.478 -1.176 -0.423 0.399 2.023 0.507 -0.115 |
| MEM | 0.052 -0.213 -1.176 0.188 0.685 0.268 1.115 |
| MIA | 0.386 -0.397 0.003 0.423 1.277 0.351 0.949 |
| MIL | -0.431 -0.354 0.040 1.596 2.076 0.712 0.830 |
| MIN | -0.688 -0.859 0.249 0.356 -0.080 -0.644 0.584 |
| NOP | -0.632 -0.713 0.302 0.154 0.560 -0.143 0.076 |
| NYK | -0.523 -0.941 -0.771 -0.901 -0.071 0.410 0.122 |
| OKC | 0.818 -0.015 0.284 1.039 1.368 -0.712 -0.623 |
| ORL | -0.355 -1.054 -1.032 0.566 0.649 -0.812 -0.719 |
| PHI | -2.155 -1.113 0.489 1.074 1.251 0.903 0.843 |
| PHO | -1.076 -1.293 -1.227 -0.686 0.713 1.075 1.601 |
| POR | 0.184 -0.352 0.346 1.284 0.793 0.508 -0.436 |
| SAC | -0.454 -0.815 -0.859 0.493 0.567 -0.141 -0.260 |
| SAS | 1.691 0.778 0.234 0.970 0.689 -0.024 -0.050 |
| TOR | 0.827 0.107 0.880 1.479 1.939 -0.418 0.696 |
| UTA | -0.053 0.183 0.330 1.085 1.338 1.150 0.732 |

### 4.3 *New model with the attendance data*

In the previous model, we only used past game histories to predict results. However, what other variable(s) could affect the predictions? We thought about the pandemic: perhaps the lack of audiences was a big factor since it was unique to the 2020-2021 season. This also explained the odd almost-random prediction above. We created an "AttdLevel" column on three levels: Full (or almost full), Partial, and Empty, according to the attendance data.

Now that we created a new model which included two pandemic seasons (19-20, 20-21) and two regular seasons (18-19, 21-22), we ran a generalized linear model to see how attendance level actually affected each team and its performance.

**Table 6 Prediction of strength coefficients for two attendance levels**

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| AttdLevelFull | 0.2498531 | 0.08870753 | 2.816594 | 0.004853579 |
| AttdLevelPartial | 0.3851477 | 0.12695576 | 3.033716 | 0.002415617 |

By the linear model (summary excerpt above), for all three attendance levels, 1 unit change in AttdLevelFull resulted in about a 0.250 change in the WinLose prediction, and 1 unit change in AttdLevelPartial resulted in a 0.385 change in the WinLose prediction. An empty arena caused no effect on the WinLose prediction.

The model showed a smaller intercept value by adding another variable to the regression. By observing the p-value for AttdLevelFull and AttdLevelPartial, both smaller than 0.000, we determined that the model was statistically significant and arrived at the conclusion that attendance level did affect the game outcome. It might seem the coefficient of AttdLevelFull was less than that of AttdLevelFul, but the difference was less than one standard error, so it wasn't a big difference. As a result, either a home arena was full or partial would not affect the outcome that much. It was more important whether there was an audience or not.

### 4.4 *Plots both on raw data set and the model predictions*

Primarily, we created a pie chart to show the distribution of the away and home games over the seven years.

Home or Away pie chart

Figure 5 Pie chart of the distribution of home and away games over seven years

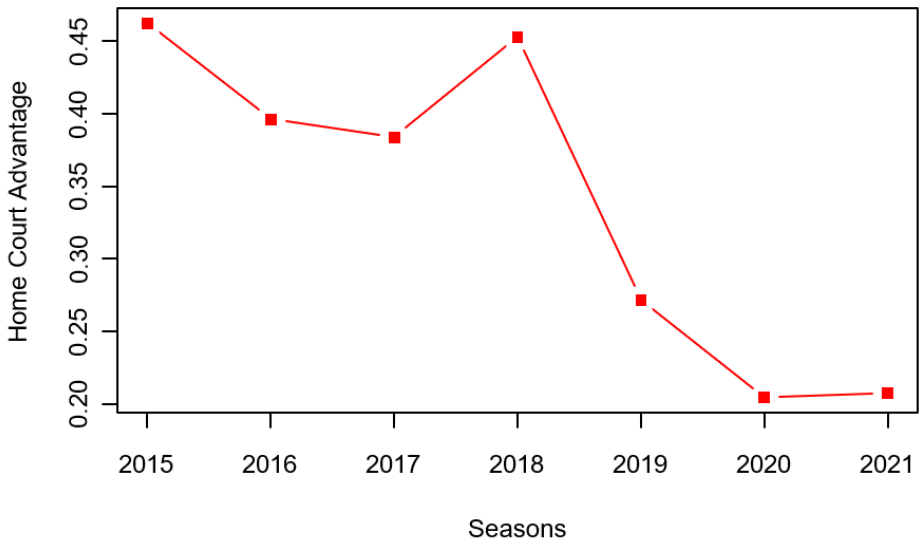


**Figure 6 Home Court Advantages Over 7 Seasons**

We created a scatter and line plot to show the home intercept and to see how the home court advantage changed over seasons.

According to the plot, there existed a clear slump during the pandemic season, which implied that the loss of audiences significantly affected the home court advantage, which plummeted from about 0.43 in previous seasons to approximately 0.20.

Finally, to help better illustrate the strength coefficients, we created two plots for the five best teams and five bottom teams over seven seasons based on their average strength coefficients. We also included WAS in both plots as the standard level for team strength.

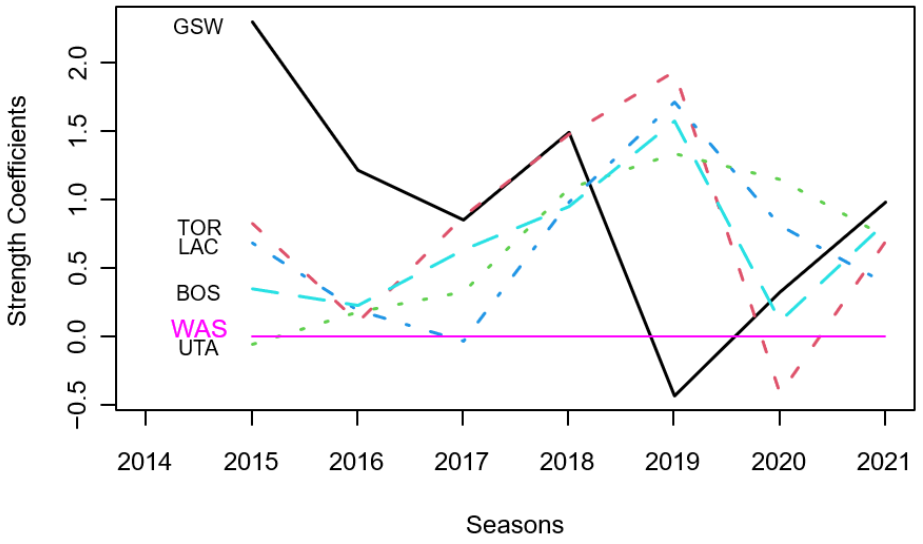


**Figure 7 Best 5 Teams Over Seasons w.r.t. WAS**

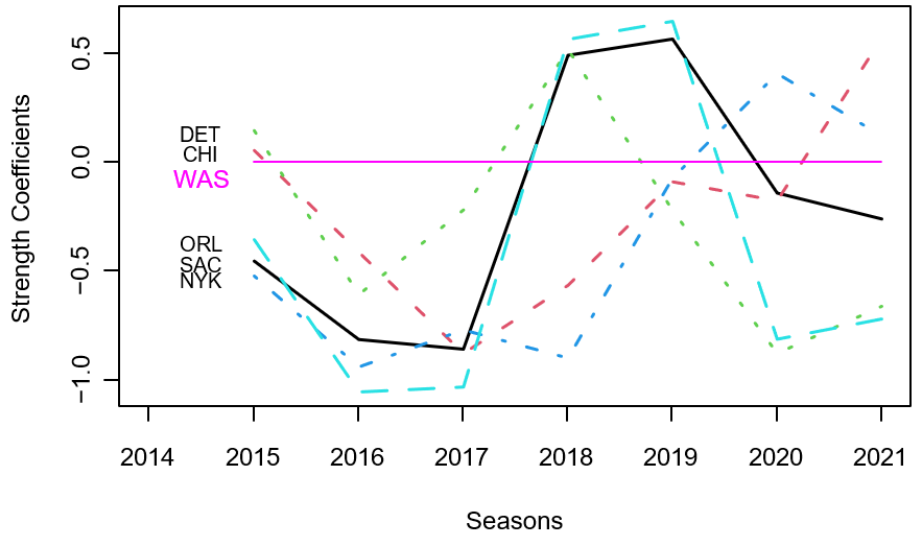


**Figure 8 Bottom 5 Teams Over Seasons w.r.t. WAS**

## 5. Conclusion

Our results revealed that there was a slight advantage towards the home team when audiences were present, but the home court
advantage was less obvious on team performance when there were partial or no audiences during the pandemic. According to the result of the strength coefficient chart, the “Home Intercept”, which showed the effect of audiences on home teams, was
significantly lower for the 2020-2021 pandemic season than those in normal seasons. When testing the accuracy of model predictions, we realized the prediction was less accurate (almost a random guess) when there were no audiences at all since there would be no “home court advantage” in games.

In the end, the study concluded that an arena with a full or partial audience would give an advantage to the home team.

## 6. References

To cite R in publications use: R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[1] Wikipedia, 2024, 2020 NBA Bubble. www.en.wikipedia.org/wiki/2020_NBA_Bubble

[2] Metcalfe, C. (2013). NBA Star Power: Impact on Attendance. Fisher Digital Publications. Sport Management Undergraduate. Paper 88. St. John Fisher University.
https://fisherpub.sjf.edu/cgi/viewcontent.cgi?article=1089&context=sport_undergrad

[3] Humphreys, B. R., & Johnson, C. (2020). The effect of superstars on game attendance: evidence from the NBA. Journal of Sports Economics, 21(2), 152-175.
DOI: https://doi.org/10.1177/1527002519885441

[4] La, V. (2014). Home team advantage in the NBA: The effect of fan attendance on performance. MPRA Paper ID 54579. Dartmouth College. https://mpra.ub.uni-muenchen.de/54579/

[5] Sports Reference LLC. Basketball Reference. www.basketball-reference.com/leagues/

[6] Wikipedia, 2024, Bradley-Terry Model. www.en.wikipedia.org/wiki/Bradley%E2%80%93Terry_model

[7] Sports Reference LLC. Basketball Reference. www.basketball-reference.com/teams/WAS/2016.html