Claremont McKenna College

Fan Impact on Home Field Advantage in Baseball

Submitted to

Professor Gelman

By

Theodore Fournier

For

Senior Thesis

Spring 2023

April 24, 2023

**Introduction**

Home field advantage (HFA) is a well-known phenomenon in sports that refers to the tendency of teams to perform better when playing on their home turf compared to playing away. Despite being widely acknowledged, the underlying reasons for HFA are still debated in the scientific literature. This thesis explores the factors contributing to HFA and its impact on athletic performance.

Research has shown that HFA exists across various sports and is a consistent phenomenon, with home teams winning a higher percentage of matches than away teams (Nevill et al., 1999). In fact, a meta-analysis of 20 studies conducted across different sports found that HFA was present in 75% of the cases (Courneya & Carron, 1992).

While several theories have been proposed to explain the reasons behind HFA, the most widely accepted one is the psychological advantage conferred by playing on one's home field. Research suggests that familiar surroundings, support from home fans, and reduced travel stress are all factors that can boost an athlete's confidence and motivation, leading to improved performance (Courneya & Carron, 1992; Neave & Wolfson, 2003).

Due to the COVID-19 pandemic, there are opportunities to create natural experiments to test effects like home fan support and last licks, the ability to go second in baseball. This is because, for the entire 2020 MLB season, no fans were in the stands, creating explicit scenarios in which fans had no impact on the impact of the game. Another result of the pandemic is that some teams were not able to play in their home ballparks, leading to situations where the away team was playing in their stadium, allowing for the testing of last-lick effects.

**Literature Review1**

There is a vast amount of literature on home field advantage and how it manifests in different sports. One of the ways home field advantage is categorized is (i) territoriality/psychological reasons, (ii) location familiarity, (iii) referee bias, (iv) crowd support, and (v) other physical reasons (Fischer & Haucap, 2020).

Due to the COVID-19 pandemic, researchers had a unique opportunity to test what impact crowd support had on home field advantage. For example, in American football, it was found that crowd support could make up as much as half of the home field advantage (Krieger & Davis 2022). By comparison, in baseball, the crowd support was found to have no statistical impact on home field advantage (Losak & Sabel 2021).

With all that has been stated above, the purpose of this paper is to explore if the results found in Losak and Sabel’s report still hold true with the inclusion of the 2021 and 2022 MLB seasons. If the results are supported, then we can safely conclude that in the case of baseball, fans have no statistical significance and the idea of "last licks" has a severe impact on home field advantage.

**Data**

The data is a combination of the 2019 through 2022 MLB season and were sourced from Baseball-Reference.com and Stathead. The data consisted of joining the schedule to the respective batting information. Once the data was joined to test to make sure that data collection was sourced and formatted correctly, a subset of the data, the 2019 and 2020 seasons, were separated and evaluated to make sure that it matched or was acceptably like Losak and Sabel’s outputs. It is expected that there would be some difference in how the data was collected, but the data returned the same statistical significance for the regressions that were run. It should be noted that *TravelDiffij* and *GameLoadDiffij* were not sourced for the data due to complexity and lack of statistical significance (Losak & Sabel 2021).

**Table 1. Summary Statistics**

(Full Data: *N* = 8,127)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Mean** | **SD** | **Min** | **Max** | **Variable** | **Mean** | **SD** | **Min** | **Max** |
| ***Weight of Home Wins*** | | | | | ***Not Home Designated*** | | | | |
| Full | 0.537 |  |  |  | Full |  |  |  |  |
| 2019 | 0.530 |  |  |  | 2019 |  |  |  |  |
| 2020 | 0.558 |  |  |  | 2020 | 0.032 |  |  |  |
| 2021 | 0.538 |  |  |  | 2021 |  |  |  |  |
| 2022 | 0.534 |  |  |  | 2022 |  |  |  |  |
| ***OPS Home*** | | | | | ***OPS Away*** | | | | |
| Full | 0.716 | 0.072 | 0.493 | 0.994 | Full | 0.714 | 0.072 | 0.482 | 0.980 |
| 2019 | 0.743 | 0.073 | 0.530 | 0.958 | 2019 | 0.738 | 0.073 | 0.535 | 0.952 |
| 2020 | 0.724 | 0.069 | 0.542 | 0.918 | 2020 | 0.720 | 0.073 | 0.544 | 0.948 |
| 2021 | 0.712 | 0.067 | 0.495 | 0.994 | 2021 | 0.711 | 0.065 | 0.482 | 0.980 |
| 2022 | 0.692 | 0.068 | 0.493 | 0.934 | 2022 | 0.690 | 0.067 | 0.491 | 0.933 |
| ***GS Home*** | | | | | ***GS Away*** | | | | |
| Full | 4.431 | 1.532 | 0.400 | 13.400 | Full | 4.505 | 1.506 | 0.800 | 11.600 |
| 2019 | 4.686 | 1.581 | 0.800 | 11.600 | 2019 | 4.713 | 1.554 | 1.000 | 11.400 |
| 2020 | 4.398 | 1.530 | 0.800 | 12.200 | 2020 | 4.772 | 1.563 | 1.000 | 11.600 |
| 2021 | 4.400 | 1.485 | 0.400 | 10.800 | 2021 | 4.477 | 1.482 | 0.800 | 9.800 |
| 2022 | 4.219 | 1.491 | 0.600 | 13.400 | 2022 | 4.225 | 1.409 | 0.800 | 9.600 |

**Methodology**

First, we identify elements that impact home field advantage using the following model:

The outcome variable, *HomeWinij,* is the outcome of a game *i* in season *j* from the home teams’ perspective (1 for a win and 0 for a loss). The model was used in both a linear and logistic regression. The sample included 2,415 regular season games from the 2019 season as the control group and all the 886 regular season games from 2020 as the treatment group, which is comparable to the 3,292 games found in Losak and Sabel (2021). The data then also contains 2,411 games from the 2021 season and 2,415 from the 2022 season. These two seasons act as a comparison to pre-COVID-19 control of 2019, with 2021 containing possible bleed-over effects and the 2022 season being the most current season, making it the most removed from the pandemic's effects. These seasons are indicated by the binary variables *Season2020j*, *Season2021j*, and *Season2022j*, respectively. The *GameDayij* variable identifies the number of games prior to and including *i* that the home team played (in 2020, most teams played 60 games in the season compared to the standard 162). This measure should be strongly correlated with the individual cumulative experiences of both the home and away teams. We interact this with the *Seasonj* variables to measure player adjustments to changes in HFA throughout a season. The *NoHomeDesignij* variable identifies if the home-designated team was playing in the away team’s ballpark.

We also included two measures of team ability; one was measured on a team level, and the other measured how effective the pitcher was. The *OPSij* and *OppOPSij* variables considered the recent on-base plus slugging average (OPS) performance of the starting lineup, and their purpose is to give a sense of how to effect each team is on the offensive side of the ball, i.e., when they are up to bat. This was calculated by taking a moving average of the last fifteen games for each team and should reflect any changes within the lineup and their impact on the game. The other measure is *GSij* and *OppGSij*, and their purpose is to measure how effective a pitcher is at preventing the other team from scoring. This statistic is heavily correlated with the winning percentage and is calculated by taking the average score from the last five games.

**Results?**

The first thing that was done was running the equation above through linear regression, the reason this was done is due to the fact a linear model is more intuitive when examining the results that are provided. The reason for being more intuitive is that the results provided are probabilities. This created the following output:

**Table 2. Home Field Advantage Linear Model**

|  |  |
| --- | --- |
| Linear Regression; Dependent Variable: Home Win (1/0) | |
| ***No Home Designated*** | -0.217\*  (0.087) |
| ***Season 2020*** | -0.046  (0.036) |
| ***Season 2021*** | 0.008  (0.027) |
| ***Season 2022*** | -0.017  (0.027) |
| ***Gameday*** | -0.000  (0.000) |
| ***X Season 2020*** | 0.002  (0.001) |
| ***X Season 2021*** | -0.000  (0.000) |
| ***X Season 2022*** | 0.000  (0.000) |
| ***Home OPS*** | 0.603\*\*\*  (0.076) |
| ***Away OPS*** | -0.591\*\*\*  (0.076) |
| ***Home GS*** | -0.087\*\*\*  (0.003) |
| ***Away GS*** | 0.078\*\*\*  (0.004) |
| Constant | 0.558\*\*\*  (0.079) |
| *N* | 8,127 |
| AIC | 10,216 |
| *Notes:* Statistical significance is defined at the \* 10%, \*\* 5%, and \*\*\* 1% levels. | |

The first thing to examine is the fact that our constant of 55.8%, the result means not only is there a 55.8% chance of winning, but the result is highly statistically significant, demonstrating the presence of HFA. This means the rest of the results throughout the paper are occurring to a data set in which HFA is present. The next thing to notice is that *NoHomeDesigntedij*, is negative and statistically significant (-21.7%), lending to the idea that one advantage of home field is being able to get the last lick in the game. So far none of the *Seasonj* measurements are statistically significant, meaning according to the data there is no statistical difference between each season. Home OPS is highly statistically significant and positive, which is to be expected, the value is essentially saying that the better the home team is at bat, the higher their chances of winning. It should be noted that Home OPS result has a higher absolute value than the Away OPS, so even if the two teams have identical OPS the home team still has an edge (1.2%) when it comes to winning. The GS variables are extremely statistically significant just like the OPS values, but it should be noted that the Away GS again has a lower magnitude than OPS but whereas before this was good for the home team, now if teams have the same GS going into a game the home team loses an edge of (0.9%).

The next step was to run the same equation through a logistic regression model with the hope that this would be more statistically useful, even at the expense of readability. The reason the model would be a better fit is due to the fact that the dependent variable, Home Win, is binary and logistic regression is better suited for the output created. The model created the following output:

**Table 3. Home Field Advantage Logistic Model**

|  |  |
| --- | --- |
| Logistic Regression; Dependent Variable: Home Win (1/0) | |
| ***No Home Designated*** | -1.034\*  (0.431) |
| ***Season 2020*** | -0.215  (0.176) |
| ***Season 2021*** | 0.043  (0.129) |
| ***Season 2022*** | -0.078  (0.130) |
| ***Gameday*** | 0.000  (0.001) |
| ***X Season 2020*** | 0.001  (0.004) |
| ***X Season 2021*** | -0.000  (0.001) |
| ***X Season 2022*** | 0.001  (0.001) |
| ***Home OPS*** | 2.925\*\*\*  (0.375) |
| ***Away OPS*** | -2.891\*\*\*  (0.373) |
| ***Home GS*** | -0.429\*\*\*  (0.018) |
| ***Away GS*** | 0.389\*\*\*  (0.019) |
| Constant | 0.275  (0.388) |
| *N* | 8,127 |
| AIC | 9,672 |
| *Notes:* Statistical significance is defined at the \* 10%, \*\* 5%, and \*\*\* 1% levels. | |

The first thing to note between the two models is that AIC was reduced from 10,216 in the linear model to 9,672 in the logistic model, what this means is that the logistic model does a better job of fitting data. Another difference between the two models is that constant is no longer statistically significant, which is expected due to the shape of a logistic regression. Also, because by reducing the impact of the constant on the regression the other variables are more likely to be more accurately weighted. We also observe that the OPS and GS statistics continue to be highly statistically significant, lending more credence to the idea that these are the primary drivers of a team’s performance at home. Most importantly for what we want to measure from the data, *NoHomeDesignatedij* remains statistically significant and none of the *seasonij* measurements were statistically significant.

**Conclusion**

Final results and what I think

**Endnotes**

1I would like to acknowledge the comprehensive and recent paper on home field advantage that Losak and Sabel (2021) accomplished. They are the foundation of this paper and I would be remiss not to give them the credit that is due.

**References**

Courneya, K. S., & Carron, A. V. (1992). The home advantage in sport competitions: A

literature review. *Journal of Sport and Exercise Psychology*, 14(1), 13-27.

Fischer, K., & Haucap, J (2020). Does crowd support drive the home field advantage in

professional soccer? Evidence from German ghost games during the COVID-19

pandemic. *DICE Discussion Paper, No. 344,* ISBN 978-3-86304-343-8,

Düsseldorf Institute for Competition Economics (DICE), Düsseldorf.

http://hdl.handle.net/10419.222278

Krieger, K. & Davis, J. L. (2022). How much of “Home Field Advantage” Comes from

the Fans? A Natural Experiment from the COVID-19 Pandemic. *UNLV Gaming*

*Research & Review Journal*, 26(5), 59-72

Losak, J. M. & Sabel, J. (2021). Baseball Home Field Advantage Without Fans in the

Stands. *International Journal of Sports Finance*, 16, 148-162.

Nevill, A. M., Balmer, N. J., & Williams, A. M. (1999). The influence of crowd noise

and experience upon refereeing decisions in football. *Psychology of Sport and Exercise*,1(3), 261-269.

Neave, N., & Wolfson, S. (2003). Testosterone, territoriality, and the ‘home advantage’.

*Physiology & Behavior*, 78(2), 269-275.