Do supporters help the home team win the match?

CPES 2 - Example of research project

Detailed instructions in Lecture 16 (https://louissirugue.github.io/metrics_on_R/lecture16/slides.html)
Grading system available here (https://louissirugue.github.io/metrics_on_R/project/grading.html)

Louis Sirugue

Spring 2023

I. Introduction

It is well established that playing home at football grants an advantage over the other team (Pollard, 1986). Yet, the specific determinants of this advantage, and the extent of their respective contributions, has not been clearly identified so far. The presence of local supporters in the stadium ranks among the most plausible reasons for this stylized fact, but whether or not supporters do help the team that plays home to win a football match is a difficult question to answer given the small variation in the presence of supporters at football matches. But by preventing supporters to attend football matches, the COVID-19 pandemic provides a unique setting to study this question. I study the effect of the presence of supporters on the probability to win the match by comparing the outcome of matches with supporters before the pandemic to those without supporters after the pandemic.

Dowie (1982) was the first to elicit a home advantage at football. Even though no causal effect could be identified, he stressed three potential reasons: fatigue for the away team due to travel, familiarity with the environment for the home team, and fans that support the home team and may play on their motivation. Evidence for these three different channels were then put forward in later studies. Concerning fatigue, Pollard et al. (2008) showed that distance traveled by the away team significantly increases the number of expected goals in favor of the home team by 0.115 goal per thousand kilometers traveled. Loughead et al. (2003) found mixed evidence about the familiarity hypothesis: high quality teams suffered after a move from their familiar venue, whereas low quality teams seemed to benefit from it. But overall, their results provide little support for facility familiarity as an explanation for the home advantage. Finally, Greer (1983) showed that booing from the crowd at basketball games had a positive effect on performances of the home team and negative effects for the team playing away. Still, the overall effect of supporters on the outcome of sports events remains to be quantified.

II. Data cleaning

I use data on every match of Premier League, Ligue 1, La Liga, and Bundesliga from season 2018-2019 to season 2020-2021. The data is publicly available at fbref.com (https://fbref.com/), and documents not only the score but also when and where the match took place, as well as the number of supporters attending the match. Each of the variables of the dataset is briefly described below.

```
# Load necessary packages
library(tidyverse)  # To manipulate the data
library(stargazer)  # To display regression results
library(kableExtra) # To make html tables

# Import data from csv file
data_match <- read.csv("data/data_match.csv")

# Display the name of each variable
names(data_match)
```

```
##  [1] "Wk"           "Day"          "Date"         "Time"         "Home"
##  [6] "xG"           "Score"        "xG.1"         "Away"         "Attendance"
## [11] "Venue"        "Referee"      "Match.Report" "Notes"        "League"
## [16] "Season"
```

The dataset contains 16 variables:

- `Wk` : Season week when the match took place
- `Day` : Week day when the match took place
- `Date` : Date of the match
- `Time` : Time of the match
- `Home` : Team that played home
- `xG` : Expected number of goals for *home* team
- `Score` : Score of the match
- `xG.1` : Expected number of goals for *away* team
- `Away` : Team that played away
- `Attendance` : Number of supporters in the stadium
- `Venue` : Name of the stadium where the match took place
- `Referee` : Name of the referee
- `Match.Report` : Link to an online report of the match
- `Notes` : Miscellaneous information on the match
- `League` : Name of the league
- `Season` : Season from 2018-2019 to 2020-2021

Not all these variables are going to be useful, so I only keep the date and time at which the match took place, the teams involved and the score, the number of supporters in the stadium, the league and the season. The following table displays the first five observations of the data.

```
data_match <- data_match %>%
  # Keep only the variables listed below in data_match
  select(Day, Date, Time, Home, Score, Away, Attendance, League, Season)

# Display the first five observations of the data
kable(head(data_match, n = 5), caption = "Outlook of the data:")
```

Outlook of the data:

Day	Date	Time	Home	Score	Away	Attendance	League	Season
Fri	2018-08-10	20:45	Marseille	4-0	Toulouse	60756	Ligue 1	2018-2019
Sat	2018-08-11	17:00	Nantes	1-3	Monaco	32760	Ligue 1	2018-2019
Sat	2018-08-11	20:00	Montpellier	1-2	Dijon	12765	Ligue 1	2018-2019
Sat	2018-08-11	20:00	Lille	3-1	Rennes	25708	Ligue 1	2018-2019
Sat	2018-08-11	20:00	Angers	3-4	Nîmes	9534	Ligue 1	2018-2019

Before starting the analysis, some variables must be recoded for convenience. For instance, the `Score` variable is not in a practical format. It stores the number of goals scored by each team, separated with a dash. I should assign the score of each team to distinct variables, and set their class to `numeric` instead of `character`. The same type of modifications can be applied to the `Time` variable, which in currently in `character` format as `hh:mm`. To transform the time variable in a continuous variable expressed in hours, the number of minutes divided by 60 should be added to the number of hours. The following table displays the first 15 lines of the data recoded as described above.

```
data_match <- data_match %>%

  # Separate the home and away score into 2 variables
  separate(Score, c("Home", "Away"), "-") %>%

  # Convert these variables as numeric
  mutate(Home = as.numeric(Home),
         Away = as.numeric(Away),

          # Generate a variable for the outcome of the match depending on who scored the
most
         Winner = case_when(Home > Away ~ "Home",
                            Home == Away ~ "Draw",
                            Home < Away ~ "Away"),

         # Recode the Time variable as a continuous variable
         Time = as.numeric(substr(Time, 1, 2)) + as.numeric(substr(Time, 4, 5)) / 60)

# Display the first 15 rows of the recoded data
kable(head(data_match, n = 15), caption = "Recoded data:")
```

Recoded data:

Day	Date	Time	Attendance	Home	Away	League	Season	Winner
Fri	2018-08-10	20.75	60756	4	0	Ligue 1	2018-2019	Home
Sat	2018-08-11	17.00	32760	1	3	Ligue 1	2018-2019	Away
Sat	2018-08-11	20.00	12765	1	2	Ligue 1	2018-2019	Away
Sat	2018-08-11	20.00	25708	3	1	Ligue 1	2018-2019	Home
Sat	2018-08-11	20.00	9534	3	4	Ligue 1	2018-2019	Away
Sat	2018-08-11	20.00	26006	2	1	Ligue 1	2018-2019	Home
Sat	2018-08-11	20.00	21421	0	1	Ligue 1	2018-2019	Away
Sun	2018-08-12	15.00	48263	2	0	Ligue 1	2018-2019	Home

Day	Date	Time	Attendance	Home	Away	League	Season	Winner
Sun	2018-08-12	17.00	23079	0	2	Ligue 1	2018-2019	Away
Sun	2018-08-12	21.00	47289	3	0	Ligue 1	2018-2019	Home
		NA	NA	NA	NA	Ligue 1	2018-2019	NA
Fri	2018-08-17	20.75	18917	1	0	Ligue 1	2018-2019	Home
Sat	2018-08-18	17.00	19003	1	3	Ligue 1	2018-2019	Away
Sat	2018-08-18	20.00	10402	1	2	Ligue 1	2018-2019	Away
Sat	2018-08-18	20.00	19300	1	0	Ligue 1	2018-2019	Home

An important step of the data cleaning process is to handle missing values. It can be seen from the table above that between each week of competition there is an empty line with missing values. These rows can be deleted by filtering out every observation for which the `Home` variable is blank.

```
# Drop blank rows
data_match <- data_match %>% filter(Home != "")
```

To check for the presence of actual missing values in the data, the following table shows the number of missing values for each variable of the dataset.

```
# Show the number of missing values for each variable
kable(data_match %>% summarise_all(~sum(is.na(.))),
      caption = "Number of missing values per variable:")
```

Number of missing values per variable:

Day	Date	Time	Attendance	Home	Away	League	Season	Winner
0	0	0	1670	0	0	0	0	0

The only variable with missing values is `Attendance`. There are 1670 matches for which the number of supporters in the stadium is not reported. To get a better understanding of what is going on with this variable, the following table summarizes the distribution of `Attendance` with its minimum and its maximum value, its mean, and the three quartiles.

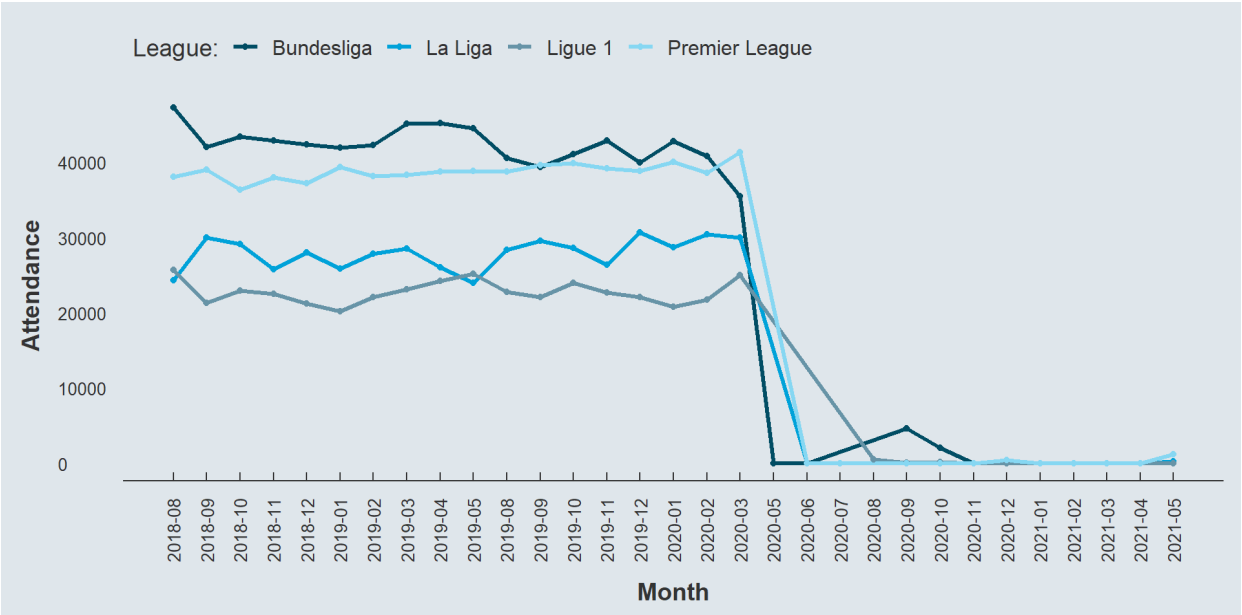
```
# Display the summary statistics of the Attendance variable
kable(as.matrix(summary(data_match$Attendance)) %>% t(),
      caption = "Attendance - Descriptive statistics:")
```

Attendance - Descriptive statistics:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
13	16158	27717	31789.99	45014	93426	1670

The number of spectators per match ranges from 13 to 93426. But due to the COVID-19 pandemic that prevented many matches from having supporters in the stadium, there should be values of `Attendance` equal to 0. It is thus possible that the missing values of `Attendance` are actually these matches where no supporter was allowed to attend the event, and that missing values are actually not missing but a way of coding no attendance. This hypothesis is even more plausible given that other than the `Attendance` variable, there is no issue of missing value in the data. A visual check can be conducted to test this hypothesis, by recoding missing values to 0 and showing the monthly evolution of the average number of supporters in stadiums. This can be done separately for each league to see whether or not the issue is league-specific.

```
attendance_data <- data_match %>%
         # Replace missing values of Attendance by 0
  mutate(Attendance = ifelse(is.na(Attendance), 0, Attendance),
         # Keep only the YYYY-MM part of the Date variable (YYYY-MM-DD)
         Month = substr(Date, 1, 7)) %>%
  # Do computations separately for each month and each league
  group_by(Month, League) %>%
  # Compute the average number of supporters in the stadium
  summarize(Attendance = mean(Attendance)) %>%
  # Sort the data by ascending order of month and group by month
  ungroup() %>% arrange(Month) %>%  group_by(Month) %>%
  # Attribute a number from 1 to N to each month whatever the league
  mutate(Month_id = cur_group_id())

ggplot(attendance_data,
       # Assign month/attendance to the x-/y-axis and one color per league
       aes(x = Month_id, y = Attendance, color = League), alpha = .75) +
  # Draw a line and a point geometry and rename legend
  geom_line(size = 1.2) + geom_point(size = 1.5) + labs(color = "League:") +
  # Label the x axis with months in character format
  scale_x_continuous(name = "Month", breaks = unique(attendance_data$Month_id),
                     labels = unique(attendance_data$Month)) +
  # Rotate the month labels by 90 degrees
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



The sudden drop to 0 attendance due to the pandemic right after March 2020 is striking, and confirms that the missing values of the `Attendance` variable should indeed be recoded as 0. It also illustrates that the COVID-19 pandemic provides an ideal setting to test for the potential effect of the presence of supporters in the stadium on the probability for the home team to win the match.

```
# Replace missing values of Attendance by 0
data_match <- data_match %>% mutate(Attendance = ifelse(is.na(Attendance), 0, Attendanc
e))
```

III. Descriptive statistics

Once the data is cleaned and recoded, it should be described with appropriate statistics. The first relevant information is the number of observations. The observation level of the data being the match, the following table displays the number of matches in the data separately for each league and each season.

```
nb_obs <- data_match %>%
  # Do computations separately for each season and each league
  group_by(League, Season) %>%
  # Compute the number of match per season/league
  summarise(n_match = n()) %>%
  # Put these values in separate columns for each season
  pivot_wider(names_from = "Season", values_from = "n_match") %>%
  # Compute the total number of matches per league
  mutate(Total = `2018-2019` + `2019-2020` + `2020-2021`)

nb_obs %>%
  # Add one Total row which is the sum of all the above
   bind_rows(nb_obs %>% mutate(League = "Total") %>% group_by(League) %>% summarise_all
(~sum(.))) %>%
  # Display in an htlm table
  kable(.,  caption = "Number of matches:") %>%
  # Set characters in the column Total in bold
  column_spec(5, bold = T) %>%
  # Set characters in the row Total in bold
  row_spec(5, bold = T)
```

Number of matches:

League	2018-2019	2019-2020	2020-2021	Total
Bundesliga	306	306	306	**918**
La Liga	380	380	380	**1140**
Ligue 1	380	279	380	**1039**
Premier League	380	380	380	**1140**
Total	**1446**	**1345**	**1446**	**4237**

The data contains a total number of 4237 observations, with slightly less observations in the 2019-2020 season than in the two others due to the cancellation of matches in Ligue 1. Besides this event, each league has 380 matches per season except the Bundesliga for which the number of matches per season amounts to 306. The following table shows the number of matches won home, away, and the number of draws, along with their respective proportion in the dataset.

```
data_match %>%
  # Do computations separately for each outcome
  group_by(Winner) %>%
  # Compute the number of observations and the percentage
  summarise(N = n(), Pct = n() / nrow(.)) %>%
  # Display in an html table
  kable(., "Distribution of match outcomes")
```

Distribution of match outcomes

Winner	N	Pct
Away	1343	0.32
Draw	1067	0.25
Home	1827	0.43

This confirm the well-established stylized fact that football matches have greater chance to be won by the home team. To provide an overview of the variables that are used in this analysis, the following tables summarizes the distribution of the three main variables: the number of supporters in the stadium, the number of goals scored by the team that plays home, and that scored by the team that plays away. These statistics are provided separately for each league and each season.

```
descriptive_data <- data_match %>%
  # Put the variables of interest in long format
  pivot_longer(c(Attendance, Home, Away),
               names_to = "Variable", values_to = "Value") %>%
  # Group the data by variable of interest, season, and league
  group_by(Variable, Season, League) %>%
  # Compute the descriptive statistics
  summarise(Min = min(Value),
            Q1 = quantile(Value, 1/4),
            Median = median(Value),
            Mean = mean(Value),
            Q3 = quantile(Value, 3/4),
            Max = max(Value)) %>%
  # Ungroup the data
  ungroup()
```

2018-2019 2019-2020 2020-2021

```
descriptive_data %>%
  # Keep only the observations of the 2018-2019 season
  filter(Season == "2018-2019") %>%
  # Keep only the variables to display
  select(-c(Variable, Season)) %>%
  # Add a caption to the table
  kable(., caption = paste("Season", "2018-2019")) %>%
  # Display the name of the variable for the corresponding rows
  pack_rows("Attendance", 1, 4) %>%
  pack_rows("Goals away", 5, 8) %>%
  pack_rows("Goals home", 9, 12)
```

Season 2018-2019

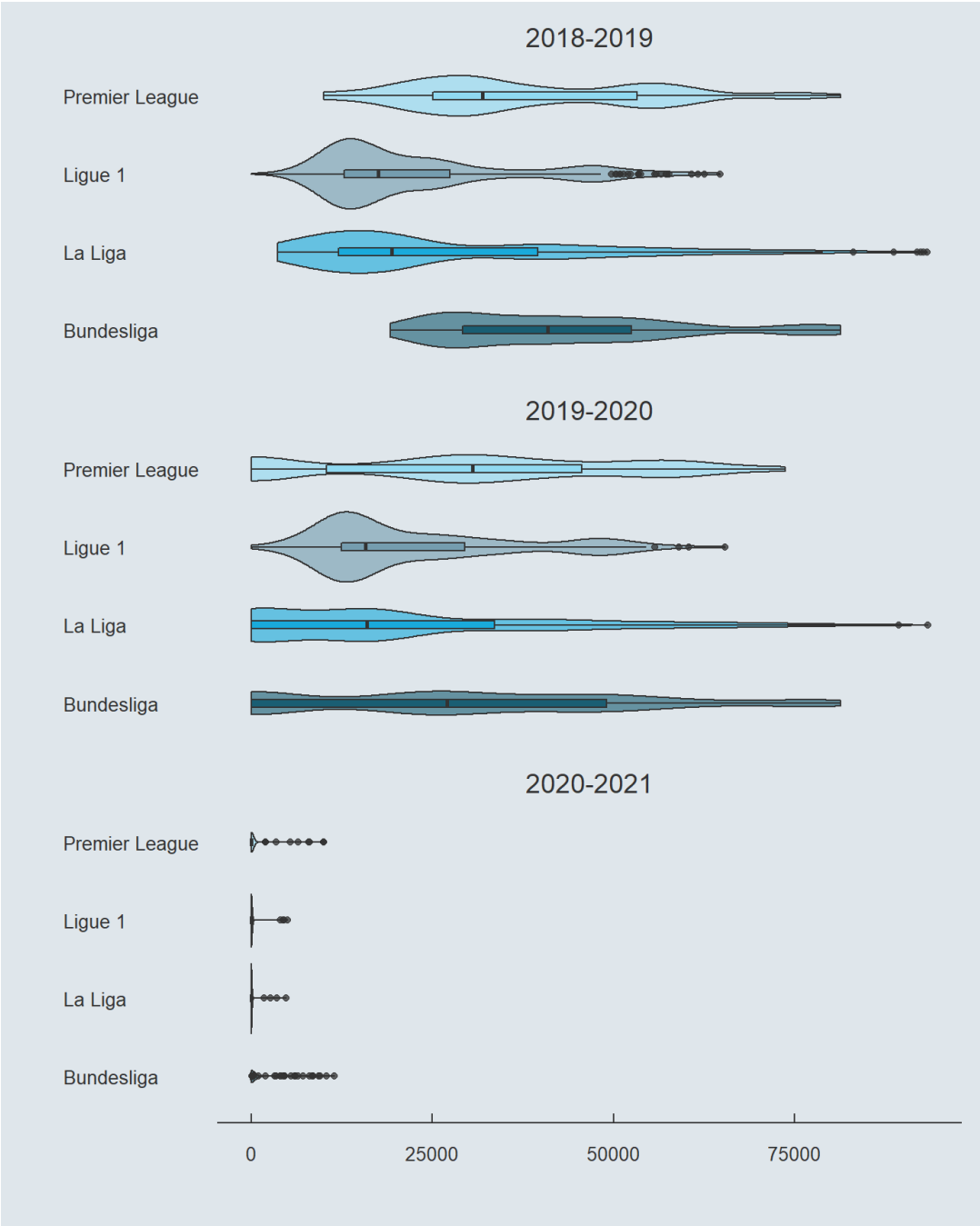
League	Min	Q1	Median	Mean	Q3	Max
Attendance						
Bundesliga	19205	29230.50	40911.0	43453.18	52500.00	81365
La Liga	3592	12074.50	19367.5	27118.68	39587.75	93265
Ligue 1	0	12795.75	17577.5	22807.27	27378.50	64696
Premier League	9980	25034.75	31948.0	38181.29	53282.75	81332
Goals away						
Bundesliga	0	0.00	1.0	1.39	2.00	6
La Liga	0	0.00	1.0	1.13	2.00	6
Ligue 1	0	0.00	1.0	1.09	2.00	5
Premier League	0	0.00	1.0	1.25	2.00	6
Goals home						
Bundesliga	0	1.00	2.0	1.79	3.00	8
La Liga	0	1.00	1.0	1.45	2.00	8
Ligue 1	0	1.00	1.0	1.47	2.00	9
Premier League	0	1.00	1.0	1.57	2.00	6

From these tables it appears that the average number of supporters in the stadium started to decline during the 2019-2020 season, to the extent that in 2020-2021 most matches in all leagues had no attendance at all, and the few matches with supporters were way below the full capacity. Also, the average and maximum number of goals scored tend to be larger for teams that play home than for teams that play away, especially for the 2018-2019 season.

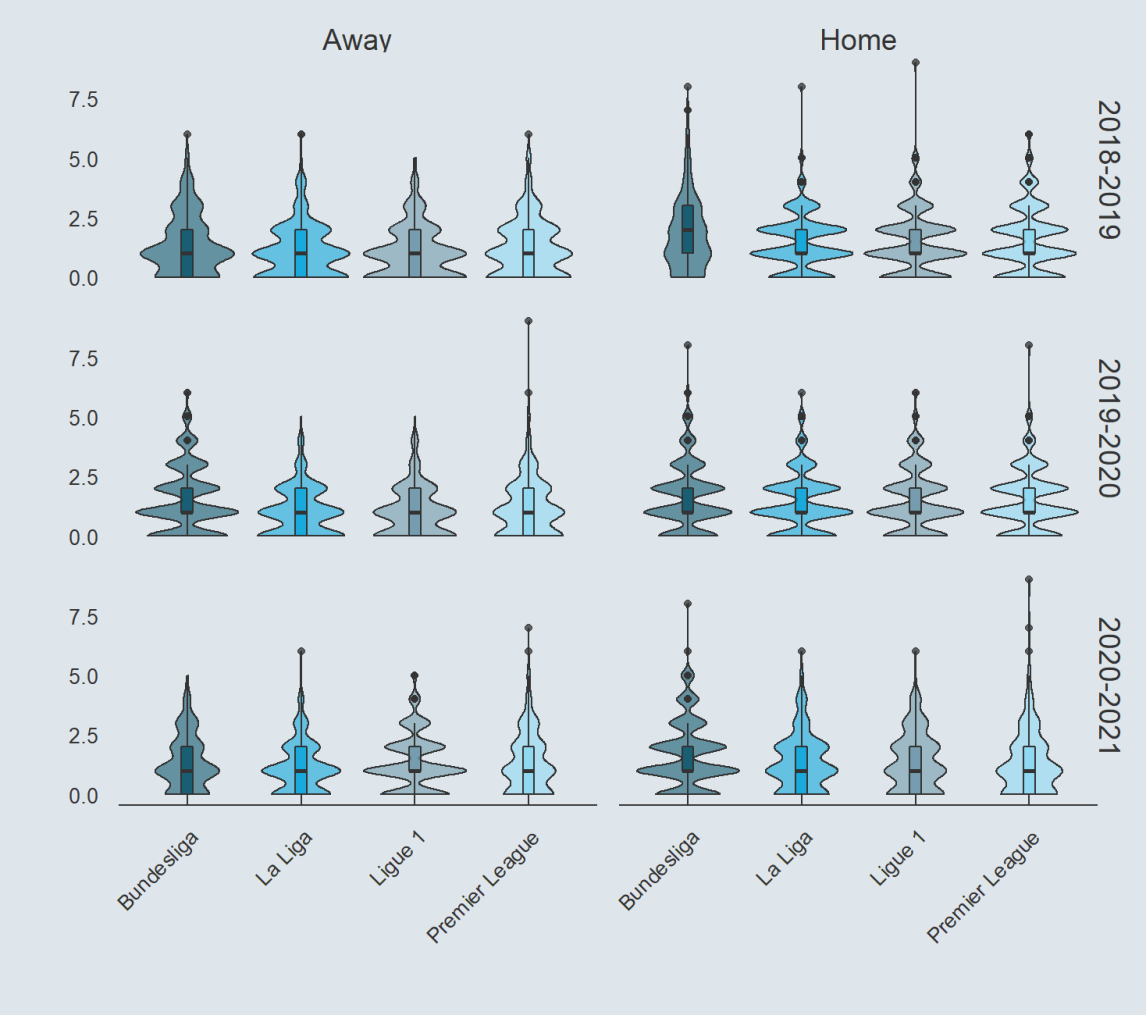
IV. Visualizing the data

To get a finer depiction of the distribution of these variables, they can be represented graphically by superimposing their density and their boxplot separately for each league and each season.

```
# Assign the League to the x and fill axes and the attendance to the y axis
ggplot(data_match, aes(x = League, y = Attendance, fill = League)) +
  # Overlay a violin density and a boxplot with transparency
  geom_violin(show.legend = F, alpha = .55) +
  geom_boxplot(width = 0.1, show.legend = F, alpha = .75) +
  # Rotate the graph and plot separately by season
  coord_flip() + facet_wrap(~ Season, ncol = 1) + ylab("") + xlab("")
```

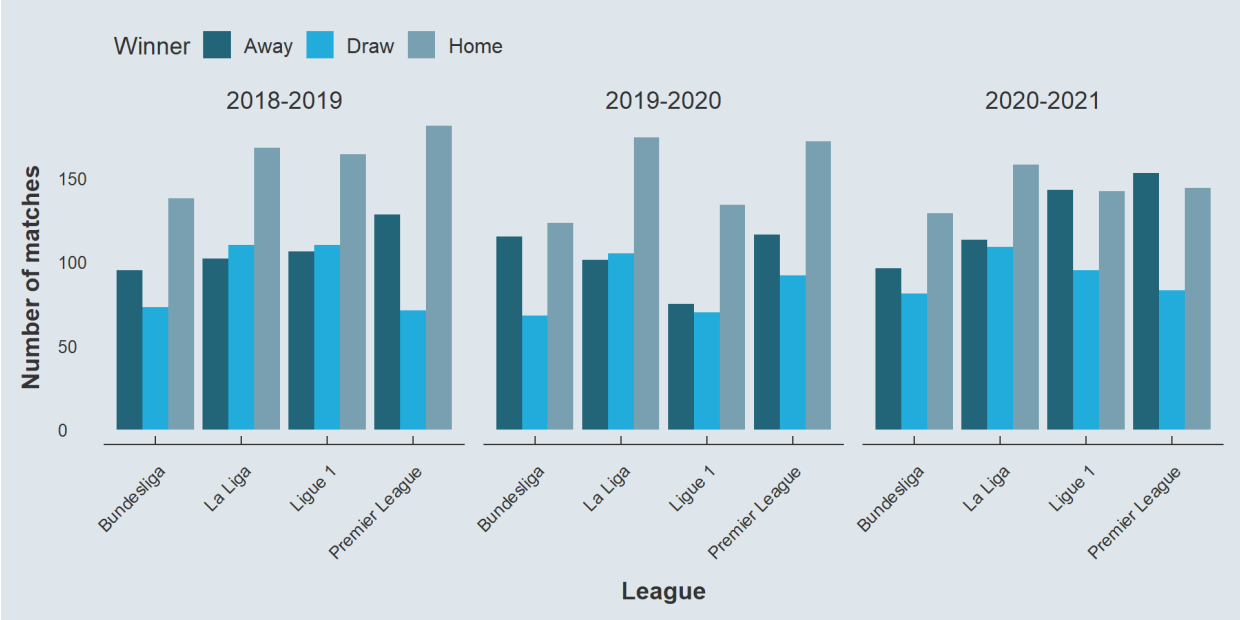


```
data_match %>%
  # Put the goals scored home and away in long format
  pivot_longer(c(Home, Away), names_to = "Variable", values_to = "Value") %>%
  # Assign the League to the x and fill axis and the goals scored to the y axis
  ggplot(., aes(x = League, y = Value, fill = League)) +
  # Overlay a violin density and a boxplot with transparency
  geom_violin(show.legend = F, alpha = .55) +
  geom_boxplot(width = 0.1, show.legend = F, alpha = .75) +
  # Plot separately by season and for home/away, and custom the axes
  facet_grid(Season ~ Variable) + ylab("") + xlab("") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1))
```



While the decline in attendance over the seasons is striking visually, it is less the case for the difference between the distributions of the goals scored home and those scored away. To get a more precise picture of the evolution of the outcome of matches over the seasons, the following graph displays the number of matches won by the home team, by the team playing away, and the number of draws, separately for each league and each season.

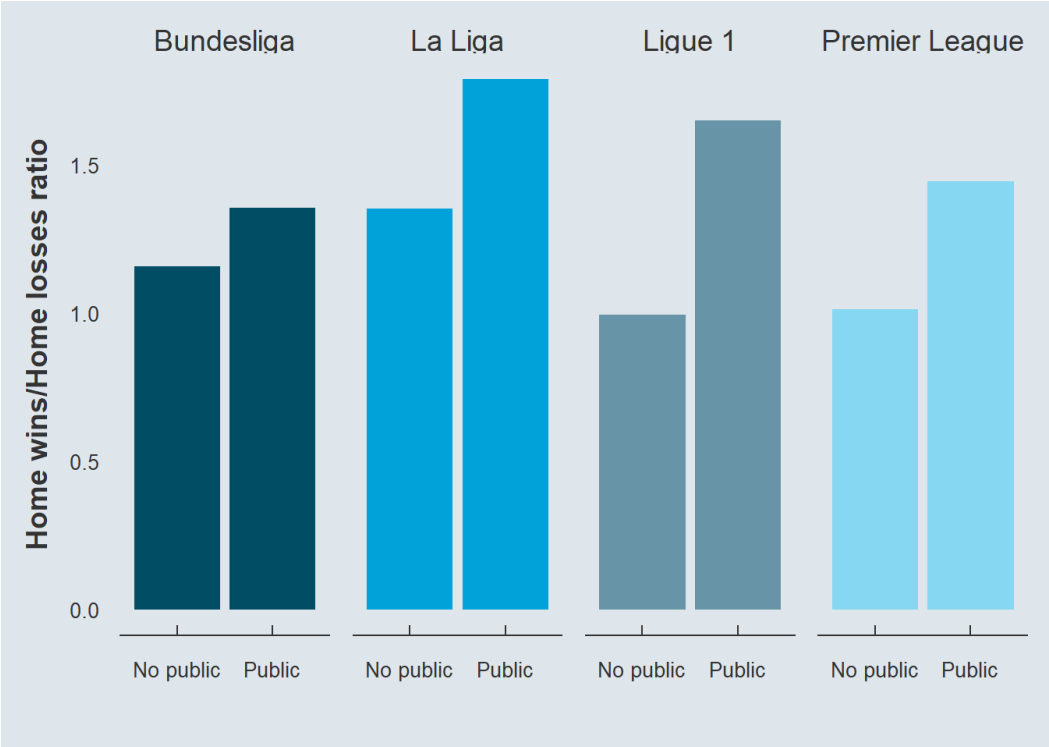
```
# Assign the League to the x axis and the outcome of the match to the fill axis
ggplot(data_match, aes(x = League, fill = Winner)) +
  # Bar plot geometry counting the number of each outcome, bars side to side
  geom_bar(stat = "count", position = "dodge", alpha = .85) +
  # Plot separately by season and custom the axes
  facet_wrap(~Season) + ylab("Number of matches") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1))
```



First, it confirms that the home team tends to win more frequently than the team that plays away. But except for the Bundesliga, it is quite clear visually that there is a decline in the difference between the number of matches won by the team that plays home and the number of matches won by the team that plays away, which seems concomitant with the restrictions imposed on attendance in stadiums.

Before estimating formally the relationship between the presence of supporters in the stadium and the probability to win the match, the following graph compares the ratio between home wins and home losses when there are supporters in the stadium and when there is none, separately for each league.

```r
# Generate a binary variable indicating the presence of supporters
data_match <- data_match %>%
  mutate(Public = ifelse(Attendance > 0, "Public", "No public"))

data_match %>%
  # Do computations separately by league and presence of supporters
  group_by(League, Public) %>%
  # Compute the ratio of home vs. away wins
  summarise(Ratio = sum(Winner == "Home") / sum(Winner == "Away")) %>%
  # Assign the presence of supporters to the x axis, the ratio to the y axis,
  # and the league to the color axis
  ggplot(., aes(x = Public, y = Ratio, fill = League), alpha = .85) +
  # Add a bar geometry to display the values side to side
  geom_bar(position = "dodge", stat = "identity", show.legend = FALSE) +
  # Plot separately for each league and custom the axes
  facet_wrap(~League, nrow = "1") + ylab("Home wins/Home losses ratio") + xlab("")
```



From this graph it is clear that the ratio between home wins and home losses tends to be higher when there are supporters in the stadium than when there is none, and that holds for the four leagues considered. But to be able to draw clear conclusions on the relationship between the presence of supporters in the stadium and the probability for the home team to win the match, a regression analysis should be carried out.

V. Regression analysis

The equation to estimate writes:

$$1\{Winner_m = \text{Home}\} = \alpha + \beta \times 1\{Public_m = \text{Yes}\} + \varepsilon_m,$$

where for a given match m the variable $1\{Winner_m = \text{Home}\}$ takes the value 1 if the winning team is that playing home and 0 otherwise, and the variable $1\{Public_m = \text{Yes}\}$ takes the value 1 if there is public in the stadium and and 0 otherwise. Because the dependent variable is binary, this equation corresponds to a linear probability model and coefficients have to be interpreted in percentage points of the probability that the home team wins the match. Because the independent variable is binary, the constant α in this model corresponds to the probability that the home team wins the match when there is no public, and the slope β corresponds to the expected percentage-point change in this probability when there are supporters in the stadium.

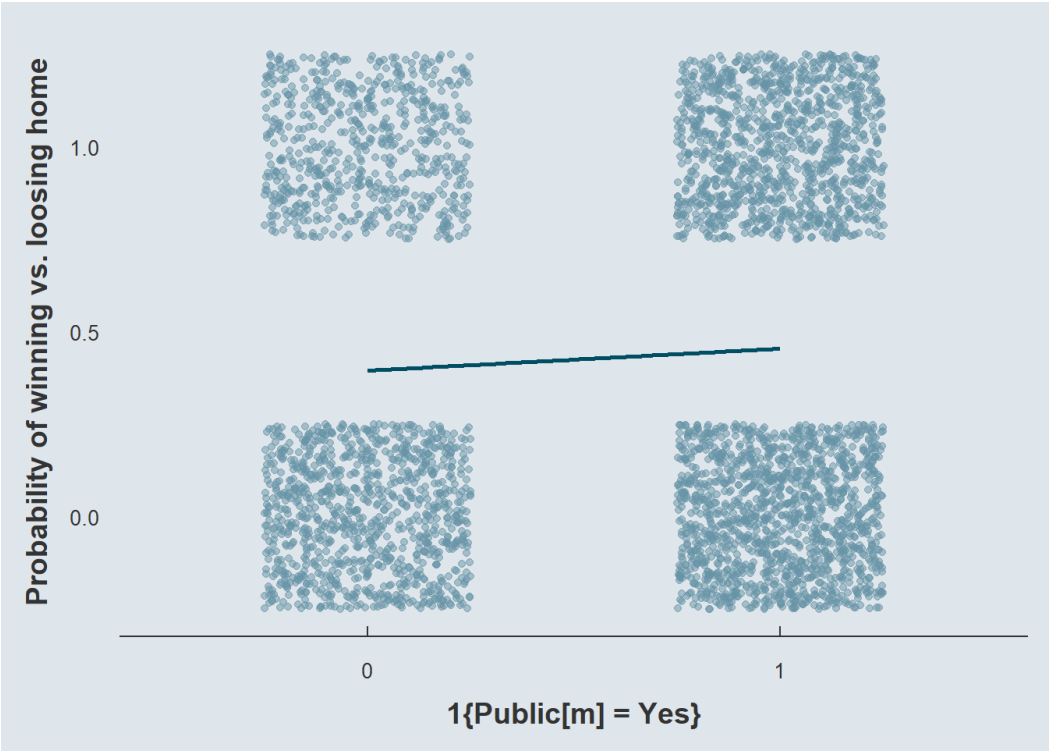
```
# Generate a binary variable that takes the value one if home team won
data_match <- data_match %>%
  mutate(Winner_home = ifelse(Winner == "Home", 1, 0))

# Estimate the regression model
stargazer(lm(Winner_home ~ Public, data_match), dep.var.labels = c("Home win"))
```

	Dependent variable:
	Home win
PublicPublic	0.059***
	(0.016)
Constant	0.395***
	(0.012)
Observations	4,237
Adjusted R^2	0.003
Note:	*p<0.1;**p<0.05;***p<0.01

According to these results, the presence of supporters in the audience increases by 5.9 percentage points on expectation the probability for the home team to win the match, everything else equal. Given that the probability for the home team to win the match (relative to loose or draw) is equal to 39.5% when there is no public, this corresponds to an average increase of about 15% in relative terms. Given that the p-values associated with $\hat{\alpha}$ and $\hat{\beta}$ are lower than 1%, these two values are statistically significantly different from 0 at the 99% confidence level.

The following plot represents the regression line estimated in the previous table. Because both the dependent and the independent variables are binary, each point can only take 4 locations on the graph. To facilitate visualization, I use `geom_jitter()` to introduce some noise in the location of each data point around these 4 possible coordinates. It appears that the number of home wins relative to the number of home losses and draws is indeed lower when there is no supporter in the stadium.

```
# Assign the dependent and the independent variables to x and y axes
ggplot(data_match, aes(x = Public, y = Winner_home)) +
  # Plot the data points with some noise to avoid overplotting
  geom_jitter(width = .25, height = .25, alpha = .5, color = "#6794A7") +
  # Plot the regression line centered with respect to the data points
  geom_smooth(data = data_match %>%
                mutate(Public = ifelse(Public == "Public", 1, 0) + 1),
              aes(x = Public, y = Winner_home),
              method = "lm", se = F, color = "#014D64") +
  # Custom the axes
  scale_x_discrete(name = "1{Public[m] = Yes}", labels = 0:1) +
  ylab("Probability of winning vs. loosing home")
```



VI. Causality assessment

The previous regression table documented a positive and statistically significant relationship between the presence of supporters in the stadium and the probability for the home team to win. These results suggest the supporters have an influence on the outcome of the match, be it directly, e.g., by impacting on the motivation of players, or indirectly, e.g., by impacting the decisions of the referees in favor of the home team.

But even though this result provides support for this hypothesis, it is not sufficient to prove the presence of a causal effect. Indeed, there may be other variables, correlated both with the dependent and the independent variable, that drive this relationship. The COVID-19 pandemic may have simultaneously prevented supporters from going to the stadium and changed the conditions for the team that plays away in a favorable way, for instance if the trip to the stadium is less tiring because there is less congestion on the roads due to remote working, or for any other reason. In other words, there may be an omitted variable bias driving part or all of the estimated relationship. Because it is not feasible to control for such variables in the regression, more sophisticated econometric specifications would be required to conclude on the causality of the effect.

VII. Robustness

But even if it is not possible to include all the relevant controls, some variables can still be added to the regression to check the robustness of the baseline result. Indeed, if the probability differential with and without supporters can be linked to changes in transport conditions with the pandemic, it could also be linked to the day in the week and the time in the day at which the match takes place, as transport conditions may also depend on that. Thus, even though controlling for these variables would not prove any irrelevance of the mechanisms mentioned in the above section, it is important to check that the baseline result is robust to the inclusion of the

variables that can be controlled for given the data available. The following table progressively includes the league, the day of the week, and the time of the day as controls in the regression. Because the `League` variables is categorical, I first set a reference category to this variable using the `relevel()` function.

```
# Set the League variable as factor and its reference category to "Premier League"
data_match <- data_match %>%
  mutate(League = relevel(as.factor(League), "Premier League"))

# Progressively include control variables in the regression
stargazer(lm(Winner_home ~ Public, data_match),
          lm(Winner_home ~ Public + League, data_match),
          lm(Winner_home ~ Public + League + Day, data_match),
          lm(Winner_home ~ Public + League + Day + Time, data_match),
          dep.var.labels = c("Home win vs. Home loss", "Home win vs. Home loss/Draw"))
```

	Dependent variable:			
	Home win vs. Home loss			
	(1)	(2)	(3)	(4)
PublicPublic	0.059***	0.060***	0.060***	0.060***
	(0.016)	(0.016)	(0.016)	(0.016)
LeagueBundesliga		-0.012	-0.012	-0.015
		(0.022)	(0.022)	(0.022)
LeagueLa Liga		0.004	0.007	-0.001
		(0.021)	(0.021)	(0.022)
LeagueLigue 1		-0.014	-0.013	-0.023
		(0.021)	(0.022)	(0.023)
DayMon			-0.039	-0.040
			(0.050)	(0.050)
DaySat			0.005	0.019
			(0.031)	(0.033)
DaySun			-0.008	0.008
			(0.031)	(0.035)
DayThu			0.006	0.009
			(0.059)	(0.059)
DayTue			0.055	0.057
			(0.047)	(0.047)
DayWed			0.023	0.026
			(0.039)	(0.039)
Time				0.004
				(0.004)
Constant	0.395***	0.400***	0.396***	0.317***
	(0.012)	(0.017)	(0.034)	(0.078)
Observations	4,237	4,237	4,237	4,237
Adjusted R^2	0.003	0.003	0.002	0.002
Note:			*p<0.1;**p<0.05;***p<0.01	

The baseline coefficient remains virtually unchanged in terms of magnitude with the inclusion of control variables, and is always statistically significantly different from 0 at 99% confidence level. Thus, the baseline estimate is robust to the inclusion of these three control variables.

Another robustness check could be performed regarding the definition of the dependent variable. Indeed, the regressions estimated so far are about the probability of winning relative to loosing or draw. An alternative definition would be to consider the probability of winning relative to loosing only, omitting draws. The following table compares the results from the baseline regression using these two possible definitions of the independent variable.

```
# Generate an outcome variable that does not account for draws
data_match <- data_match %>%
  mutate(Winner_home2 = ifelse(Winner != "Draw", Winner_home, NA))

# Regress whether the home team won on the presence of supporters for these
# two definitions of the reference group
stargazer(lm(Winner_home ~ Public, data_match),
          lm(Winner_home2 ~ Public, data_match),
          dep.var.labels = c("Home win vs. Home loss", "Home win vs. Home loss/Draw"))
```

	Dependent variable:	
	Home win vs. Home loss	Home win vs. Home loss/Draw
	(1)	(2)
PublicPublic	0.059***	0.078***
	(0.016)	(0.018)
Constant	0.395***	0.529***
	(0.012)	(0.014)
Observations	4,237	3,170
Adjusted R^2	0.003	0.006
Note:		*p<0.1;**p<0.05;***p<0.01

Using this alternative definition, it appears that even though the presence of supporters is associated with a higher probability to win for the home team, even with no public the home team is still more likely to win than the team playing away, by about 3 percentage points ($\hat{\alpha} > 50\%$). The coefficient of interest is statistically significantly different from 0 at 99% confidence level for both variable definitions. In terms of magnitude, the coefficients from the two definitions cannot be compared directly because they are mechanically inflated by the omission of the possibility of draw, but the ratio of the effect of public in the stadium on the probability to win, relative to the probability to win when there is no public, is very similar in the two cases ($\frac{0.059}{0.395} = 0.1494 \approx 0.1475 = \frac{0.078}{0.529}$). It is thus reasonable to conclude that this result is also robust to variations in the definition of the reference category of the outcome variable.

VIII. Heterogeneity

But the fact that the coefficient is robust does not mean that it is homogeneous. To investigate whether the relationship differs from one league to another, the independent variable of interest should be interacted with the `League` variable, which is equivalent to estimating the regression separately for each league.

```
# Progressively control and interact with League in the regression
stargazer(lm(Winner_home ~ Public, data_match),
          lm(Winner_home ~ Public + League, data_match),
          lm(Winner_home ~ Public + League + Public * League, data_match),
          dep.var.labels = c("Home win"))
```

	Dependent variable:		
	Home win		
	(1)	(2)	(3)
PublicPublic	0.059***	0.060***	0.074**
	(0.016)	(0.016)	(0.030)
LeagueBundesliga		-0.012	0.008
		(0.022)	(0.035)

LeagueLa Liga		0.004	0.019
		(0.021)	(0.032)
LeagueLigue 1		-0.014	-0.016
		(0.021)	(0.034)
PublicPublic:LeagueBundesliga			-0.032
			(0.045)
PublicPublic:LeagueLa Liga			-0.025
			(0.042)
PublicPublic:LeagueLigue 1			0.002
			(0.044)
Constant	0.395***	0.400***	0.392***
	(0.012)	(0.017)	(0.023)
Observations	4,237	4,237	4,237
Adjusted R^2	0.003	0.003	0.002
Note:		*p<0.1;**p<0.05;***p<0.01	

Column (3) shows that the coefficient of interest for the reference category, Premier League, amounts to 7.4 percentage points and is statistically different from 0 at the 95% confidence level. The difference between the effect in Premier League and that in other leagues range from -3.2 percentage points (i.e., an effect of 4.2 percentage points, for Bundesliga) to 0.2 percentage points (i.e., an effect of 7.6 percentage points, for Ligue 1). Yet, because the coefficients associated with the interaction terms are not significant, we cannot conclude that these different league-specific effects are statistically significant from each other. In other words, there is no evidence of a heterogeneity of the effect across leagues.

IX. Conclusion

In this analysis I use data on football matches in Premier League, Ligue 1, La Liga, and Bundesliga, from season 2018-2019 to season 2020-2021, to investigate the relationship between the presence of supporters in the stadium and the probability for the football team that plays home to win the match. The estimation of this relationship relies on the fact that the COVID-19 pandemic prevented supporters from going to the stadium, such that the outcome of these matches can be compared to those played in regular conditions. Graphical evidence indeed show a clear and sudden drop to 0 attendance in stadiums, concomitant to the pandemic right after March 2020.

Results show that the presence of supporters in the audience increases by 5.9 percentage points on expectation the probability for the home team to win the match, everything else equal. Yet, this result may not be interpreted as causal if the COVID-19 pandemic have simultaneously prevented supporters from going to the stadium and changed the conditions for the team that plays away relative to the conditions for the team that plays home. In addition, the external validity of the result is not granted, as it is estimated using four European football leagues only. Still, the estimated coefficient appears to be robust to controlling for the league, the day of the week, and the time of the day, as well as changes in the definition of the outcome variable, and results show no evidence for a heterogeneity of the effect across leagues.

References

Dowie, J. (1982). Why Spain should win the world cup. *New Scientist*, 94(10), 693-695.

Greer, D. L. (1983). Spectator booing and the home advantage: A study of social influence in the basketball arena. *Social psychology quarterly*, 252-261.

Loughead, T. M., Carron, A. V., Bray, S. R., & Kim, A. J. (2003). Facility familiarity and the home advantage in professional sports. *International Journal of Sport and Exercise Psychology*, 1(3), 264-274.

Pollard, R. (1986). Home advantage in soccer: A retrospective analysis. *Journal of sports sciences*, 4(3), 237-248.

Pollard, R., Silva, C. D., & Medeiros, N. C. (2008). Home advantage in football in Brazil: differences between teams and the effects of distance traveled. *Revista Brasileira de Futebol (The Brazilian Journal of Soccer Science)*, 1(1), 3-10.