Does the presence of supporters make football games more interesting?

Intro to econometrics using R - Example of research project

Adam Soliman

```r
library(tidyverse)
library(stargazer)
library(readxl)
```

Introduction

As any football fan knows, the outcome of a game is affected by many factors. While some of them are observable, such as team budget, players' quality, or team ranking, others, such as team cohesion or luck, are much more difficult to measure. Moreover, some determinants of a game's results, such as the home advantage, are well established, yet the relative importance of the various factors is still to be determined.

In particular, the role of supporters on football results is still largely unclear. Researchers have argued that the presence of local supporters in a stadium was one of the channels of the home advantage. They have also shown that large numbers of supporters can influence refereeing. But do supporters also galvanise players and push them to "play harder"? More precisely, the research question we will investigate is: is the total number of goals scored in a game affected by the number of supporters present in the stadium?

Literature review

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.

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: (i) fatigue for the away team due to travel, (ii) familiarity with the environment for the home team, and (iii) 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.

Data

To answer this question, we will be using data on every match of Ligue 1 from season 2014-2015 to season 2018-2019. We do not include any later year as we do not want to include the period where attendance to games was impossible during the Covid-19 pandemic. The data is publicly available at fbref.com, a website which collects all sorts of statistics and information about football games and players. For every game scheduled in these seasons, the data includes the score as well as when and where the match took place and the number of supporters attending the match.

While this data is very detailed, it lacks some variables that could be important, most importantly a measure of team quality. To complement it, we will use data on budgets of every team in the Ligue 1 from each season. This measure, which we obtain from the https://www.sportune.fr/sport-business/asse-om-psg-tous-les-budgets-de-la-ligue-1-en-2014-2015-103736/2 will be used as a proxy of the team quality. The data from Sportune are obtained are the basis of numbers computed by the French sports daily L'Équipe.

Another important factor that could be related to both score and game attendance is the weather, but gathering daily weather data from 2014 to 2019 was deemed too much work.

```
#uploading all yearly files, keeping relevant variables and creating a variable for the season
ligue1_18_19 <- read.csv("Ligue1_18_19.csv", encoding="UTF-8") %>%
    select(Wk, Day, Date, Time, Home, Score,
           Away, Attendance, Venue, Referee) %>%
    mutate(Season="2018-2019")

ligue1_17_18 <- read.csv("Ligue1_17_18.csv", encoding="UTF-8") %>%
    select(Wk, Day, Date, Time, Home, Score,
           Away, Attendance, Venue, Referee) %>%
    mutate(Season="2017-2018")

ligue1_16_17 <- read.csv("Ligue1_16_17.csv", encoding="UTF-8") %>%
    select(Wk, Day, Date, Time, Home, Score,
           Away, Attendance, Venue, Referee) %>%
    mutate(Season="2016-2017")

ligue1_15_16 <- read.csv("Ligue1_15_16.csv", encoding="UTF-8") %>%
    select(Wk, Day, Date, Time, Home, Score,
           Away, Attendance, Venue, Referee) %>%
    mutate(Season="2015-2016")

ligue1_14_15 <- read.csv("Ligue1_14_15.csv", encoding="UTF-8") %>%
    select(Wk, Day, Date, Time, Home, Score,
           Away, Attendance, Venue, Referee) %>%
    mutate(Season="2014-2015")

#bind them all into a single file and remove yearly files
ligue1 <- rbind(ligue1_14_15, ligue1_15_16, ligue1_16_17, ligue1_17_18, ligue1_18_19)
rm(ligue1_14_15, ligue1_15_16, ligue1_16_17, ligue1_17_18, ligue1_18_19)


#looking inside my data
head(ligue1)
```

```
##   Wk Day       Date  Time     Home Score          Away Attendance
## 1  1 Fri 2014-08-08 20:30    Reims   2-2     Paris S-G      18540
## 2  1 Sat 2014-08-09 21:00     Nice   3-2      Toulouse      19474
## 3  1 Sat 2014-08-09 21:00 Guingamp   0-2 Saint-Étienne      16852
## 4  1 Sat 2014-08-09 21:00   Nantes   1-0          Lens      27964
## 5  1 Sat 2014-08-09 21:00    Evian   0-3          Caen       9915
## 6  1 Sat 2014-08-09 21:00    Lille   0-0          Metz      34327
##                                       Venue           Referee    Season
## 1                Stade Auguste-Delaune II   Stéphane Lannoy 2014-2015
## 2                           Stade de Nice  Franck Schneider 2014-2015
## 3                      Stade du Roudourou     Mikael Lesage 2014-2015
## 4 Stade de la Beaujoire - Louis Fonteneau Sébastien Desiage 2014-2015
## 5                         Parc des Sports      Saïd Ennjimi 2014-2015
## 6                     Stade Pierre-Mauroy     Frédy Fautrel 2014-2015
```

```
#checking that the position of the dash in the score is always the same
ligue1 %>%
    mutate(position=str_locate(Score, "-")) %>%
    group_by(position) %>%
    summarise(n=n())
```

```
## # A tibble: 1 x 2
##   position[,"start"] [,"end"]     n
##                <int>    <int> <int>
## 1                  2        2  1906
```

```
#creating variables for goals by team and winner
ligue1_clean <- ligue1 %>%
    mutate(goals_home = as.integer(substr(Score, 1, 1)),
           goals_away = as.integer(substr(Score, 3, 3)),
           goals_total = goals_home + goals_away,
           winner_type=case_when(goals_away>goals_home ~ "Away",
                            goals_away<goals_home ~ "Home",
                            goals_away==goals_home ~ "Draw"),
           winner_team=case_when(goals_away>goals_home ~ Away,
                            goals_away<goals_home ~ Home,
                            goals_away==goals_home ~ "Draw"),
           Date=as.Date(Date), #convert date
           Paris_Home=Home=="Paris S-G", #variables for Paris
           Paris_Away=Away=="Paris S-G",
           Paris_Played = Paris_Home | Paris_Away,
           Month = format(Date, "%m")) %>%
    select(-Score) %>%
    filter(!is.na(Wk)) #drop matches outside of the regular competition weeks


#read budget data
budgets <- read_excel("budgets_ligue1.xlsx") %>%
    mutate(Budget=str_replace(Budget, "M€", ""), #Remove the unit at the end of the line
           Budget=str_replace(Budget, ",", ".")) #replace decimal comma by decimal point

#check whether all budgets could be converted to numeric
budgets %>%
    filter(is.na(as.numeric(Budget))) %>%
    select(Budget)
```

```
## Warning: There was 1 warning in `filter()`.
## i In argument: `is.na(as.numeric(Budget))`.
## Caused by warning:
## ! NAs introduced by coercion
## # A tibble: 1 x 1
##   Budget
##   <chr>
## 1 35-40
```

```
#fix the single non-numeric budget, replacing by average
budgets <- budgets %>%
    mutate(Budget=ifelse(Budget=="35-40", "37.5", Budget),
           Budget=as.numeric(Budget))

#create a budget ranking by season

budgets <- budgets  %>%
    group_by(Saison) %>%
    mutate(Rank_Budget= rank(desc(Budget)))

#Merge to have columns with budget of home and away teams
ligue1_budget <- left_join(ligue1_clean, budgets,
                          by=c("Home"="Club", "Season"="Saison")) %>%
    rename(Budget_Home=Budget,
           Rank_Budget_Home=Rank_Budget) %>%
    left_join(budgets,
              by=c("Away"="Club", "Season"="Saison")) %>%
    rename(Budget_Away=Budget,
           Rank_Budget_Away=Rank_Budget) %>%
    mutate(Budget_Dif = abs(Budget_Home - Budget_Away),
           Budget_Sum = Budget_Home + Budget_Away,
           Top3_Played = Rank_Budget_Away<=3 | Rank_Budget_Home<=3)
```

Once all yearly files are bound together, the dataset contains information on 1900 games. For 9 of them attendance is missing even though there is a score, indicating that the game did happen. A quick search online of these games helps us understand that these missing values actually correspond to games that were played with no spectators.

```
nrow(ligue1_budget)
```

```
## [1] 1900
```

```
ligue1_budget %>%
    filter(!is.na(goals_total) & is.na(Attendance)) %>%
    nrow()
```

```
## [1] 9
```

```
ligue1_budget <- ligue1_budget %>%
    mutate(Attendance = ifelse(is.na(Attendance), 0, Attendance))

ligue1_budget %>% select(goals_total) %>% summary()
```

```
##   goals_total
##  Min.   :0.000
##  1st Qu.:1.000
##  Median :2.000
##  Mean   :2.581
##  3rd Qu.:4.000
##  Max.   :9.000
```

Descriptive statistics

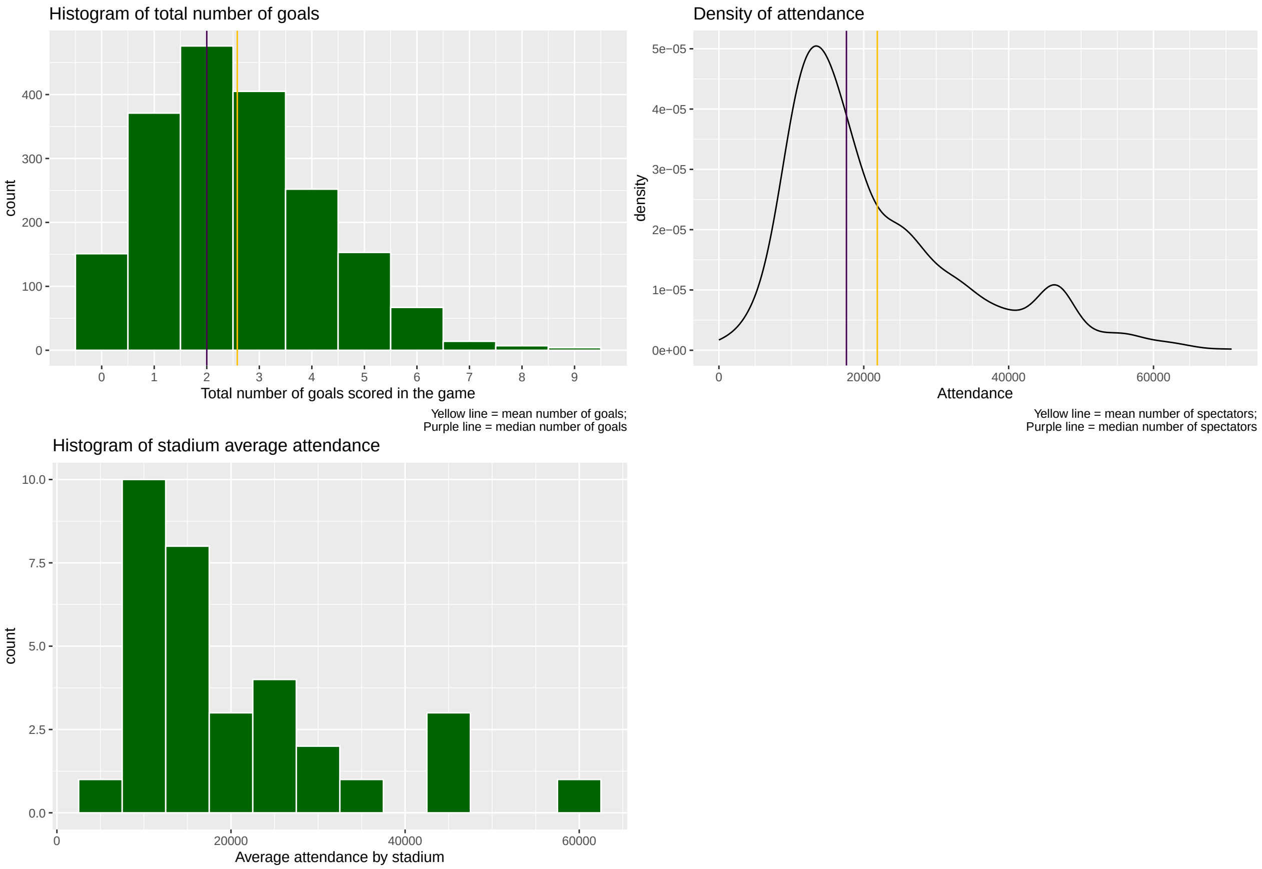
The total number of goals in those 1900 games varies between 0 and 9, with an average of around 2.6. The histogram included below shows that the distribution of goals is slightly right skewed: there are few games where more than 5 goals are scored.

The attendance variable has an interesting distribution, as we can see from the density plot below. While most games have between 10,000 and 20,000 spectators, the distribution is very right skewed, with many games welcoming more than 40,000 fans. In particular we can see that there is a bump in the distribution between 45,000 and 50,000 spectators. This bump is likely due to the fact that not all stadiums have the same capacity. While most cities in France would have similar-sized stadiums, the largest cities like Paris, Lyon, Marseille or Lille would be able to welcome more fans in their stadiums. This should inform our later analysis, since if those cities also have higher quality teams we could be mistaking the effect of the presence of supporters for the effect of the home team being from a large city. To understand how much the venue could affect the number of spectators in a game, we plot an histogram of the average attendance by stadium, which shows that indeed some stadiums welcome on average less than 10,000 people, while others welcome on average over 60,000 fans.

```r
ligue1_budget %>%
    summarise(`Mean goals scored`=mean(goals_total),
              `Median goals scored` = median(goals_total),
              `Mean attendance`=mean(Attendance),
              `Median attendance`=median(Attendance),
              `Min attendance`=min(Attendance),
              `Max attendance`=max(Attendance))
```

```
##   Mean goals scored Median goals scored Mean attendance Median attendance
## 1          2.580526                   2         21867.8           17604.5
##   Min attendance Max attendance
## 1              0          70785
```

```r
ligue1_budget %>%
    ggplot() +
    aes(x=goals_total) +
    geom_histogram(binwidth=1, boundary=0.5, color="white", fill="darkgreen") +
    scale_x_continuous(breaks=0:9) +
    geom_vline(xintercept = mean(ligue1_budget$goals_total), color="#ffc000") +
    geom_vline(xintercept = median(ligue1_budget$goals_total), color="#440154FF") +
    labs(title = "Histogram of total number of goals",
    caption="Yellow line = mean number of goals;
    Purple line = median number of goals",
    x="Total number of goals scored in the game")


ligue1_budget %>%
    ggplot() +
    aes(x=Attendance) +
    geom_density() +
    geom_vline(xintercept = mean(ligue1_budget$Attendance), color="#ffc000") +
    geom_vline(xintercept = median(ligue1_budget$Attendance), color="#440154FF") +
    labs(title = "Density of attendance",
    caption="Yellow line = mean number of spectators;
         Purple line = median number of spectators") +
    scale_x_continuous(minor_breaks=seq(0, 70000, 5000))

ligue1_budget %>%
    group_by(Venue) %>%
    summarise(mean_attendance = mean(Attendance)) %>%
    ggplot() +
    aes(x=mean_attendance) +
    geom_histogram(binwidth=5000, boundary=2500, color="white", fill="darkgreen") +
    labs(title = "Histogram of stadium average attendance",
         x="Average attendance by stadium") +
    scale_x_continuous(minor_breaks=seq(0, 70000, 5000))
```

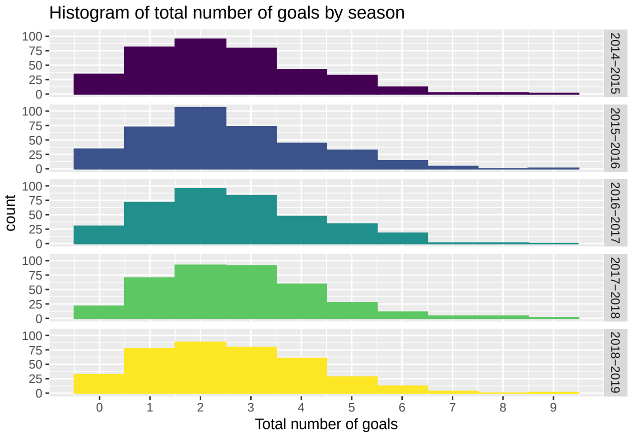


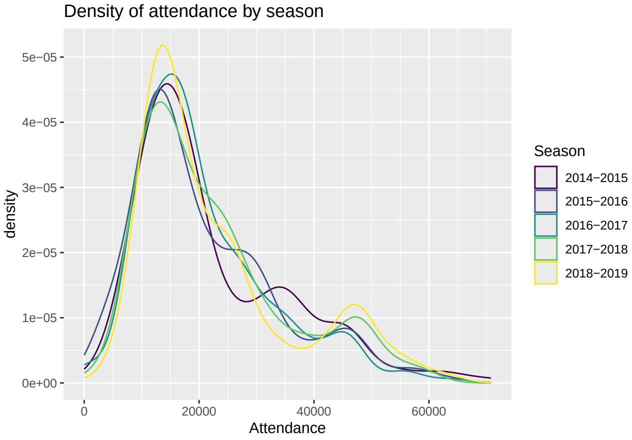
The same graph and statistics by year show us that there was no clear change across time in the distribution of goals scored or number of spectators.

```
ligue1_budget %>%
    group_by(Season) %>%
    summarise(`Mean goals scored`=mean(goals_total),
              `Median goals scored` = median(goals_total),
              `Mean attendance`=mean(Attendance),
              `Median attendance`=median(Attendance),
              `Min attendance`=min(Attendance),
              `Max attendance`=max(Attendance))
```

```
## # A tibble: 5 x 7
##   Season    `Mean goals scored` `Median goals scored` `Mean attendance`
##   <chr>                   <dbl>                 <dbl>             <dbl>
## 1 2014-2015                2.49                     2            22271.
## 2 2015-2016                2.53                     2            20995.
## 3 2016-2017                2.61                     2            20867.
## 4 2017-2018                 2.72                       3             22398.
## 5 2018-2019                 2.56                       2             22807.
## # i 3 more variables: `Median attendance` <dbl>, `Min attendance` <dbl>,
## #   `Max attendance` <dbl>
```

```
ligue1_budget %>%
    ggplot() +
    aes(x=goals_total, color=Season, fill=Season) +
    geom_histogram(binwidth=1, boundary=0.5, show.legend = F) +
    scale_x_continuous(breaks=0:9) +
    scale_color_viridis_d() +
    scale_fill_viridis_d() +
    labs(x = "Total number of goals",
         title = "Histogram of total number of goals by season") +
    facet_grid(rows = vars(Season))


ligue1_budget %>%
    ggplot() +
    aes(x=Attendance, color=Season) +
    geom_density() +
    scale_x_continuous(minor_breaks=seq(0, 70000, 5000))+
    scale_color_viridis_d() +
    scale_fill_viridis_d() +
    labs(title = "Density of attendance by season")
```

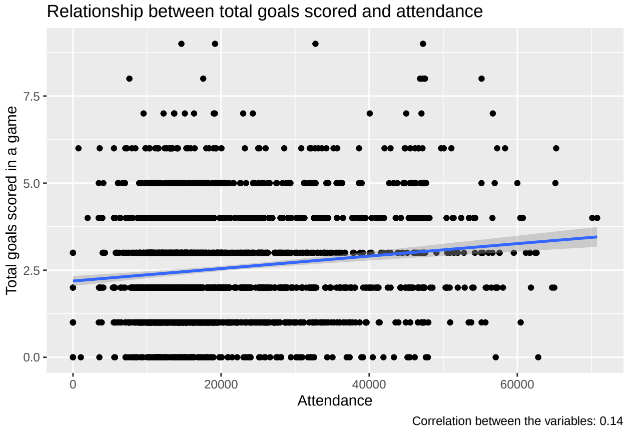




To get a first glimpse of the relationship we are interested in, we also plot the number of goals scored in a game against the number of spectators present. We can clearly see that, while there is a slight positive relationship between the two variables, it is not very strong as the points are very dispersed.

```
correl = as.character(round(cor(ligue1_budget$Attendance, ligue1_budget$goals_total), 2))

ligue1_budget %>%
    ggplot() +
    aes(x=Attendance, y=goals_total) +
    geom_point() +
    geom_smooth(method = "lm") +
    labs(y="Total goals scored in a game",
         title = "Relationship between total goals scored and attendance") +
    labs(caption = paste0("Correlation between the variables: ", correl))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Relationship between total goals scored and attendance

Correlation between the variables: 0.14

Methodology

In order to estimate the causal effect of spectators on the number of goals scored in a game, we will be using regression analysis. In a first specification, we will run a simple regression of the total number of goals on the attendance to the game. Because the attendance varies a lot, from no fan at all to over 70,000 we will use a transformed version of our variable of interest: attendance in thousands of people. The resulting regression will have the following formula, where m is a game:

$$goals_m = \beta_0 + \beta_1 \times attendance_m + \varepsilon_m$$

However, we are well aware that such a regression is not enough to uncover causality, for different reasons.

A first obvious issue of this specification is the issue of omitted variables. In order to obtain a causally identified result, we need to show that if the number of spectators changed, and nothing else changed in parallel then the number of goals would also change as a result. However, it is extremely likely that an increase in the number of spectators is correlated with other factors. If these other factors are in turn correlated with the number of goals, our estimate of β_1 will not be the true causal effect of the number of spectators: it will be biased. We can think of many such factors.

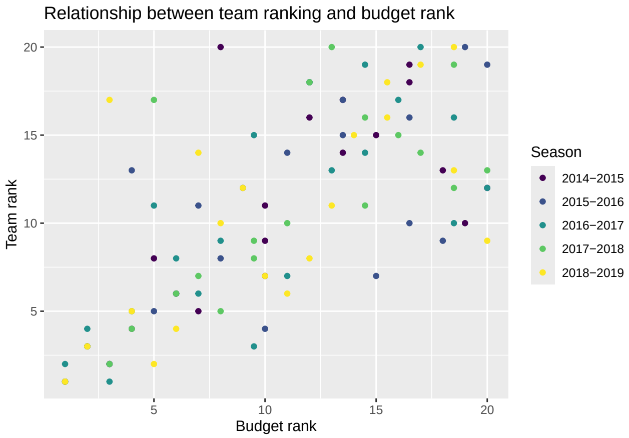
For instance, as we saw in the distribution of spectators by venue, some stadiums such as the Parc des Princes and Stade de France stadiums have a much higher capacity than most others in France. Given that the Paris Saint Germain, which plays in those stadiums, has won the ligue 1 in almost all the years we study (except the 2016-2017 season), we can safely assume that they had a higher quality team, which scored more goals than others. Larger stadiums are also located in larger cities, which would have a larger population to draw fans from (Paris has more fans than Guingamp). For these reasons, we will include the home team as a control variable in later regressions. We expect this effect of the Home team to capture both the effect of stadium size and the effect of the sphere of influence, which is why we prefer this variable to the venue where the game is held.

Another possible omitted variable is the overall quality of the team. It is possible that, even for a given stadium (and hence home team), more fans want to attend the game when the opposing team is a high quality one, to see famous players in action or to support their team in a difficult game. This result has been show before in the literature in Portugal and Germany where people come in greater numbers to see games where the away team is Bavaria Munich or Benfica, Sporting, and Porto. https://www.tandfonline.com/doi/pdf/10.1080/13504851.2011.639725 https://journals.sagepub.com/doi/full/10.1177/1527002516661602#. To control for this, we will use two alternative methods. The first one will be to control for the budget of the two teams in presence. We argue that budget is a good proxy for quality, in a context where the "price" of best football players is out of reach of average teams. We can also check that budget is an adequate proxy by computing the correlation between each team's season ranking and their budget. We can clearly see that while the correlation is not very linear, the teams with the higest budgets are in the highest ranks at the end of the season.

```
ligue1_18_19_results <- read.csv("Ligue1_results_18_19.csv", encoding="UTF-8") %>%
    select(Rk, Squad) %>%
    mutate(Season="2018-2019")

ligue1_17_18_results <- read.csv("Ligue1_results_17_18.csv", encoding="UTF-8") %>%
    select(Rk, Squad) %>%
    mutate(Season="2017-2018")

ligue1_16_17_results <- read.csv("Ligue1_results_16_17.csv", encoding="UTF-8") %>%
    select(Rk, Squad) %>%
    mutate(Season="2016-2017")

ligue1_15_16_results <- read.csv("Ligue1_results_15_16.csv", encoding="UTF-8") %>%
    select(Rk, Squad) %>%
    mutate(Season="2015-2016")

ligue1_14_15_results <- read.csv("Ligue1_results_14_15.csv", encoding="UTF-8") %>%
    select(Rk, Squad) %>%
    mutate(Season="2014-2015")

#bind them all into a single file and remove yearly files
ligue1_results <- rbind(ligue1_14_15_results, ligue1_15_16_results,
                        ligue1_16_17_results, ligue1_17_18_results,
                        ligue1_18_19_results)
rm(ligue1_14_15_results, ligue1_15_16_results, ligue1_16_17_results,
   ligue1_17_18_results, ligue1_18_19_results)


ligue1_budget_results <- left_join(ligue1_results, budgets,
                           by=c("Squad"="Club", "Season"="Saison"))

cor(ligue1_budget_results$Rank_Budget, ligue1_budget_results$Rk)
```

```
## [1] 0.7032525
```

```
ggplot(ligue1_budget_results, aes(x=Rank_Budget, y=Rk, color=Season)) +
    geom_point() +
    scale_color_viridis_d() +
    labs(title = "Relationship between team ranking and budget rank",
         x = "Budget rank",
         y = "Team rank")
```



We will control for budget in two ways: first, we will sum the budget of the two teams playing. The higher the measure is, the higher the total quality of the teams in presence. Second, we will create a dummy for having one of the top 3 budgets, as previous research has shown that in Portugal this "top-3" threshold was an explanatory factor of attendance.

Another related control we will include is a dummy for whether Paris played during the game: it is very clear that PSG was the dominating team in those 6 seasons, and it is highly likely that it drew larger crowds and also scored more goals than other teams.

Finally, we will control for other potential omitted variables, such as the month of the year when the game is played, which we argue could be correlated with attendance, because the weather differs and because the stakes of each game differ throughout the year. It's also possible that both those elements (weather and stakes) could directly affect the number of goals scored. we will also include the season when the game is played, because it is possible that the characteristic of the Ligue 1 differed by year, especially given that the players in the league are not constant across time.

We will include each of those control variables independently, to see how they affect our point estimate of interest, before we run a final regression which includes all of our control variables.

Results and analysis

```r
reg_simple <- lm(goals_total ~ I(Attendance/1000), ligue1_budget)

stargazer(reg_simple, type="latex",
          dep.var.labels="Total goals scored",
          covariate.labels="Attendance in thousands",
          keep.stat=c("n", "rsq", "adj.rsq"))
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Fri, Aug 23, 2024 - 07:23:40

The results of this simple regression can be interpreted as follows: on average, an increase in attendance by one thousand spectators is associated with an increase in the total number of goals scored of 0.018 goals. We can see that this coefficient is significant at the 1% level: we can reject the null hypothesis that the coefficient of interest is equal to zero at the significance level 0.01.

Even if it was causal, this number would be pretty low: to increase the number of goals scored in a game by 1 goal would require increasing the number of fans in the stadium by 55,556 (1/0.018). This is huge, particularly since the highest attendance recorded is 70,000!

Lastly, the R2 of the regression is 0.02 which means 2% of the variance in goals scored is explained by attendance. This implies attendance is not a particularly predictive factor of the total number of goals scored.

Table 1:

	Dependent variable:
	Total goals scored
Attendance in thousands	0.018***
	(0.003)
Constant	2.189***
	(0.073)
Observations	1,900
R^2	0.020
Adjusted R^2	0.020
Note:	*p<0.1; **p<0.05; ***p<0.01

But as discussed previously, it is likely that this regression suffers from omitted variable bias, or even reverse causality. To try to disentangle the effect of attendance from that of other variables, we will progressively include control variables in our regression.

```r
reg_season <- lm(goals_total ~ I(Attendance/1000) + Season, ligue1_budget)
reg_month <- lm(goals_total ~ I(Attendance/1000) + Month, ligue1_budget)

stargazer(reg_season, reg_month, type="latex",
          dep.var.labels="Total goals scored",
          covariate.labels="Attendance in thousands",
          keep.stat=c("n", "rsq", "adj.rsq"))
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Fri, Aug 23, 2024 - 07:23:41

In table 2, we include two control variables alternatively: the season in which the game is played, and the month of the year. The first thing to notice is that the point estimate we are interested in does not change. This means that on average an increase in attendance by one thousand spectators is associated with an increase in the total number of goals scored of 0.018 (0.017) goals, keeping the season (resp. month) constant. This implies that not including those variables in the regression does not create a large omitted variable bias. This could occur if the variables are not highly correlated with attendance, or with the number of goals scored. The coefficient of interest also keeps its significance level.

```r
reg_home <- lm(goals_total ~ I(Attendance/1000) + Home, ligue1_budget)

stargazer(reg_home, type="latex",
          dep.var.labels="Total goals scored",
          covariate.labels="Attendance in thousands",
          font.size = "scriptsize",
          no.space = T,
          keep.stat=c("n", "rsq", "adj.rsq"))
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Fri, Aug 23, 2024 - 07:23:41

The way to interpret the coefficient on attendance in table 3 is as follows: for a given home team, on average, increasing attendance by one thousand spectators is associated with a 0.024 goals increase in the total number of goals scored. The result is still significant at the 1% level. The fact that the point estimate increased compared to the simple regression implies that omitting the Home team variable was creating a negative

Table 2:

	Dependent variable:	
	Total goals scored	
	(1)	(2)
Attendance in thousands	0.018***	0.017***
	(0.003)	(0.003)
Season2015-2016	0.057	
	(0.117)	
Season2016-2017	0.141	
	(0.117)	
Season2017-2018	0.224*	
	(0.116)	
Season2018-2019	0.056	
	(0.116)	
Month02		0.034
		(0.160)
Month03		0.235
		(0.171)
Month04		0.304*
		(0.158)
Month05		0.309*
		(0.172)
Month08		0.019
		(0.168)
Month09		0.079
		(0.162)
Month10		−0.038
		(0.171)
Month11		0.088
		(0.167)
Month12		0.160
		(0.164)
Constant	2.093***	2.079***
	(0.104)	(0.132)
Observations	1,900	1,900
R^2	0.022	0.025
Adjusted R^2	0.020	0.020
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$	

Table 3:

	Dependent variable:
	Total goals scored
Attendance in thousands	0.024^{***}
	(0.006)
HomeAngers	−0.066
	(0.316)
HomeBastia	−0.372
	(0.333)
HomeBordeaux	−0.071
	(0.317)
HomeCaen	−0.027
	(0.308)
HomeDijon	0.446
	(0.333)
HomeEvian	0.073
	(0.446)
HomeGazélec Ajaccio	0.764^{*}
	(0.448)
HomeGuingamp	0.182
	(0.306)
HomeLens	−0.457
	(0.449)
HomeLille	−0.485
	(0.337)
HomeLorient	0.026
	(0.333)
HomeLyon	0.052
	(0.366)
HomeMarseille	−0.327
	(0.386)
HomeMetz	0.579^{*}
	(0.335)
HomeMonaco	0.595^{*}
	(0.305)
HomeMontpellier	0.170
	(0.305)
HomeNancy	−0.148
	(0.449)
HomeNantes	−0.451
	(0.319)
HomeNice	−0.083
	(0.312)
HomeNîmes	0.253
	(0.447)
HomeParis S-G	0.465
	(0.384)
HomeReims	0.130
	(0.333)
HomeRennes	−0.238
	(0.315)
HomeSaint-Étienne	−0.296
	(0.328)
HomeStrasbourg	0.259
	(0.376)
HomeToulouse	0.120
	(0.307)
HomeTroyes	−0.017
	(0.365)
Constant	2.041^{***}
	(0.266)
Observations	1,900
R^2	0.055
Adjusted R^2	0.041
Note:	*p<0.1; **p<0.05; ***p<0.01

omitted variable bias. This means that on average, home teams which attract a lower attendance tend to host games where more goals are scored. A possible reason for this could be team quality: bad teams are likely to attract less supporters, and they would also hold games where more goals are scored (by their opponents!).

To account for this, we will now turn to the control variables which we decided to use as a proxy for quality: sum of budget, whether a top 3 team played, and whether Paris played.

```
reg_budgets_sum <- lm(goals_total ~ I(Attendance/1000) + Budget_Sum, ligue1_budget)
reg_top_3 <- lm(goals_total ~ I(Attendance/1000) + Top3_Played, ligue1_budget)
reg_Paris <- lm(goals_total ~ I(Attendance/1000) + Paris_Played, ligue1_budget)
reg_all <- lm(goals_total ~ I(Attendance/1000) + Budget_Sum + Top3_Played + Paris_Played, ligue1_budget

stargazer(reg_budgets_sum, reg_top_3, reg_Paris,reg_all,
          type="latex",
          dep.var.labels="Total goals scored",
          covariate.labels=c("Attendance in thousands", "Sum of teams' budget in millions", "A top 3  bu
          keep.stat=c("n", "rsq", "adj.rsq"))
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Fri, Aug 23, 2024 - 07:23:41

Table 4:

	Dependent variable:			
	Total goals scored			
	(1)	(2)	(3)	(4)
Attendance in thousands	0.006^{*}	0.011^{***}	0.012^{***}	0.008^{**}
	(0.003)	(0.003)	(0.003)	(0.004)
Sum of teams' budget in millions	0.002^{***}			0.001
	(0.0003)			(0.001)
A top 3 budget team played		0.597^{***}		0.401^{**}
		(0.085)		(0.162)
Paris played			0.587^{***}	-0.072
			(0.133)	(0.296)
Constant	2.132^{***}	2.162^{***}	2.251^{***}	2.134^{***}
	(0.073)	(0.072)	(0.074)	(0.088)
Observations	1,900	1,900	1,900	1,900
R^2	0.041	0.045	0.030	0.046
Adjusted R^2	0.040	0.044	0.029	0.044
Note:		*p<0.1; **p<0.05; ***p<0.01		

Focusing on the first column in this table, we can see that while the coefficient on attendance is still significant (at the 10% level only), its magnitude has almost been divided by three. This suggests that not including the budget of the teams in presence created a very large and positive omitted variable bias. In other words, games where teams have a higher budget tend to draw more crowds and to see more goals scored in the end. In the first column, another thing of interest is the fact that the magnitude of the coefficient on the budget is quite small: a 1 million euro increase in the teams' budgets sum is associated, on average, with a 0.002 increase in the total number of goals.

Columns (2) and (3) show similar results. While the inclusion of dummies for whether a team with a top 3 budget played or whether Paris played has no effect on the significance level of our coefficient, it also decreases its magnitude, suggesting a similar omitted variable bias. Including all three controls in the same regression decreases the coefficient magnitude and the significance level. Interestingly, the sum of the teams' budget is no longer significant.

Finally, in order to control for all factors together, and get closer to the ceteris paribus assumption, we run a regression including all of our control variables: the season, the month, the home team, and the game quality as captured by our three alternative measures. We do not show the coefficients for Home team, because they are not directly of interest to us, and they would make the table too long

```r
reg_all <- lm(goals_total ~ I(Attendance/1000) + Budget_Sum + Top3_Played + Paris_Played +
                      Season + Month + Home, ligue1_budget)
stargazer(reg_all, type="latex",
          dep.var.labels="Total goals scored",
          covariate.labels=c("Attendance in thousands",
                             "Sum of teams' budget, in thousands",
                             "A top 3  budget team played", "Paris played"),
          no.space = T,
          keep.stat=c("n", "rsq", "adj.rsq"),
          omit=c("Home"),
          add.lines=list(c("Home team included", "Yes", "Yes", "Yes")))
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Fri, Aug 23, 2024 - 07:23:41

The results of this table are very similar to that of table 4. The first thing we notice is that including the full set of controls makes our point estimate no longer significant at the 10% level. This means that we cannot reject the null hypothesis that the β_1 we are trying to estimate is actually equal to zero. Again, it seems that whether a top 3 budget team played is actually driving most of the relationship between attendance and the number of goals.

Whether these estimates are causal is still unclear. The main weakness of our study is that we are not fully addressing the issue of reverse causality: is it the games with many spectators that see more goals, or games where more goals are expected which draw more spectators? In order to understand better the direction of the causality, we could use two methods. The first one would be to use existing literature on the determinants of sport games attendance, to try to better understand what are the determining factors and include them in our regression. Another, potentially more convincing method, would be to use the exogenous shock of the covid 19 pandemic as a variation in the number of spectators that is not driven by their expectation of the game's quality, but by a law outside of their control.

Another more general issue with our regressions is the fact that we might not be controlling for team's quality in a very accurate way. A future study would benefit from using data on players to construct a better measure of team quality, which would potentially be less noisy than our budget measure.

Something else that we are not controlling for are the stakes of each specific game: is the game a "classic duel" between two teams, is it determinant for the league victory, or for qualification to an european league? These types of special games are likely to attract more fans, and it is possible that the stakes either galvanise the players and increase the total number of goals, or make them more tense and lead to 0-0 draws.

Conclusion

While the analysis that we conducted shows that there is a significant relationship between the number of fans present in a football stadiums and the number of goals that are scored, it appears that this relationship mostly disappears when controlling for the budgets of the teams playing. It is not clear that spectators have a causal effect on goals, as their presence might simply be related to the expected quality of the players

Table 5:

	Dependent variable:
	Total goals scored
Attendance in thousands	0.012
	(0.007)
Sum of teams' budget, in thousands	0.001
	(0.001)
A top 3 budget team played	0.352*
	(0.190)
Paris played	−0.464
	(0.373)
Season2015-2016	0.028
	(0.123)
Season2016-2017	0.103
	(0.123)
Season2017-2018	0.134
	(0.129)
Season2018-2019	−0.053
	(0.133)
Month02	0.029
	(0.158)
Month03	0.222
	(0.169)
Month04	0.300*
	(0.157)
Month05	0.325*
	(0.172)
Month08	0.024
	(0.165)
Month09	0.080
	(0.160)
Month10	−0.044
	(0.169)
Month11	0.071
	(0.165)
Month12	0.157
	(0.162)
Constant	1.818***
	(0.305)
Home team included	Yes
Observations	1,900
R^2	0.073
Adjusted R^2	0.051
Note:	*p<0.1; **p<0.05; ***p<0.01

on the field, which in turns would be correlated with the number of goals. In order to better answer this question, one could use the covid shock! (But don't do it in your finals, it would be too easy)

Bibliography

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.