Home-Field Advantage and COVID-19: An Analysis of Serie A Soccer

Damary Gutierrez Hernandez, [redacted]

February 4, 2022

1. Introduction

Home field advantage is a well-known and studied phenomenon in sports. Factors such as players' comfort with the physical environment (maintaining practice and rest routines when the team does not have to travel), home team confidence and expectation to win, and home audience behavior influencing referee behavior have been proposed as explanations of home field advantage. It has been found that this phenomenon is most prevalent in football, and thus Serie A data provides a good opportunity to study home field advantage in greater detail (Jamieson 2010). Previous studies using the percentage of points earned by the home team of all points earned in all games played have found that these factors do indeed play a role in home field advantage (Pollard and Gómez 2014). However, this most widely used approach is not ideal for analysis of individual teams, and so other studies have chosen to look at goals instead of points to address such issues and gain a better understanding of home field advantage (Marek and Vávra 2020).

Previous studies have also looked at the physiological effects of fans at a home game being a "12th man" contributing to referee bias and home teams' performances. One study looked into referee favoritism of the home team in football and found that referees are more likely to display favoritism if the potential for satisfying the crowd is bigger. By looking at the amount of injury time referees add at the end of the game and "intense" situations, essentially when it is a close match and injury time is more important, the study found that if the home team is losing, referees are more likely to allow for greater injury time, therefore favoring the home team by allowing them more time to possibly score. Additionally, referees are more likely to allow for less injury time (i.e. signaling the end of a match) if the match is at a draw and then the home team scores (Garciano et al.). It is also believed that coaching and playing behavior is impacted by fan support, i.e. coaches setting more offensive tactics and having higher expectations for winning and players playing more aggressively (Petiot et al. 2021; Goumas 2012). One study looked at pre-COVID and post-COVID team performance (number of corner shots, number of shots, and number of shots on target) and referee decisions (number of fouls, number of yellow cards, and number red cards) and found that the absence of fan support post-COVID resulted in weaker home team performance and a greater penalization on home teams by referees in comparison to pre-COVID, indicating the significance of home field advantage in football matches (Bilalić et al.).

In this project, we aim to quantify home field advantage, defined by the home team performing better when it is at its home field. To do so, we will predict expected goal differential between the home and away teams while controlling for selected explanatory variables. Such explanatory variables will allow us to control for the impact of the pandemic on the home field advantage phenomenon (when little to no spectators were allowed to attend matches). Additionally, our model will look to control for other impactful factors in a football match, such as differences in skill levels and differences in the number of yellow cards, red cards, shots, shots on target, corner kicks, and shots hit woodwork.

1.1 Dataset

In this case study, we are using Serie A match data from Football-Data from 2009 to 2021 and a dataset ranking of the teams throughout the seasons. The match datasets for each season provide information on number of goals, full and half time results, number of shots, shots on target, hit woodwork, corners, red cards, and yellow cards for home and away team. The Serie A Standings dataset provides capacity, points, current season ranking, and prior season ranking for each football club. We merged the match datasets for each season with the Serie A Standings dataset to identify home team, away team, and season for each observation (each match).

1.2 Variables

Our response variable is continuous variable equal to the difference in goals between the home and away teams for a given match. This variable was derived from the variables for the number of full time goals made by the home team and the number of full time goals made by the away team.

We chose explanatory variables that would capture potential effects of the "12th man phenomenon," referee bias, and coaching/team behavior. We created a new variable that indicates an average percentage of the home stadium filled for a match derived from the capacity of the home stadium and average attendance at that stadium per season to better understand how the size of the spectators present at the game would influence home field advantage (Serie A - Attendance Figures). Due to the great amount of skew present in this attendance variable (attendance was fairly constant until COVID-19 when many stadiums were barred from having attendees), this variable was median-centered in order to make meaningful inference. We also created variables for the difference in yellow cards and the difference in red cards using the variables for home team yellow cards, away team yellow cards, home team red cards, and away team red cards to assess the potential presence of referee bias. Variables for the difference in shots, shots on target, corner kicks, and times hit woodwork, respectively, were also created to control for one team playing more aggressively or offensively than another. In order to account for the differences in skill level for the participating football clubs, we also created a variable for the difference in team ranks at the time in which two clubs play.

1.3 Exploratory Data Analysis

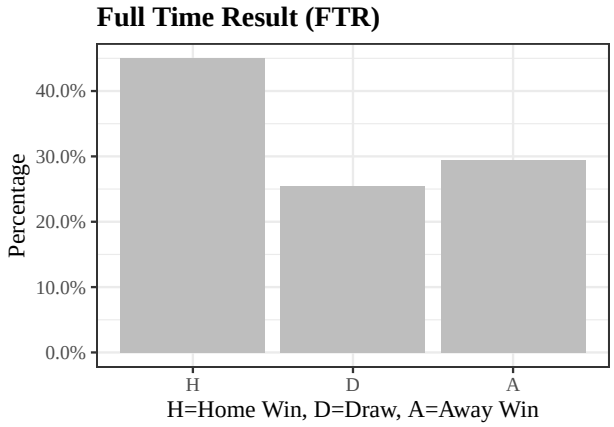

Figure 1: Full Time Result and Win Plurality

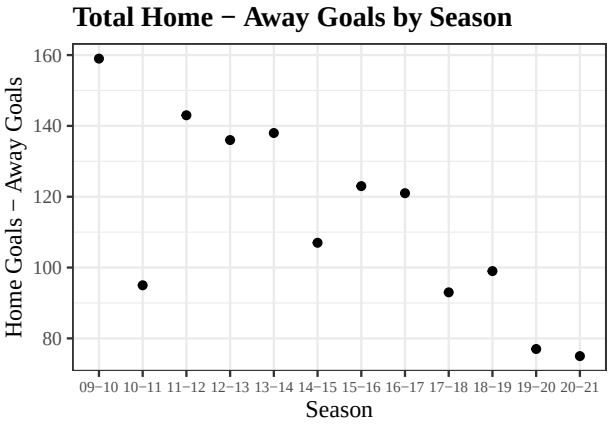

Figure 2: Goal Differential by Season

Figure 1 shows that the plurality of Serie A games from 2009-2021 result in a win for the home team. Figure 2 shows the total annual goal differential between home and away teams in the same time span. As expected, home teams always score more goals than away teams in a season, but the goal differential varies greatly from season to season.

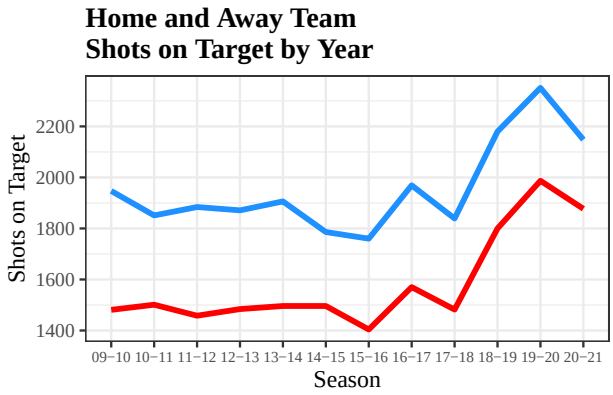

Figure 3

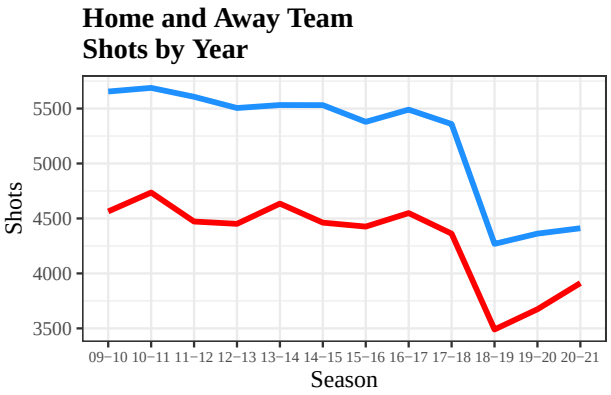

Figure 4

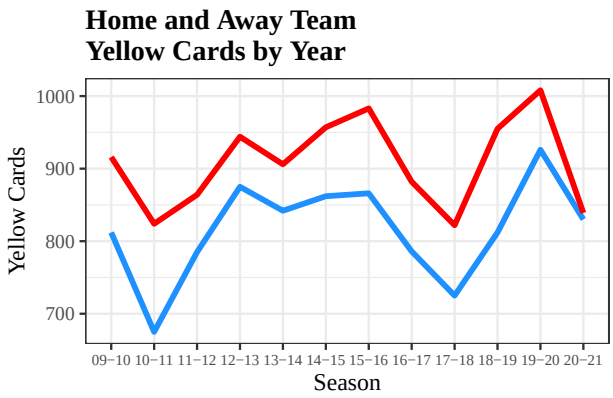

Figure 5

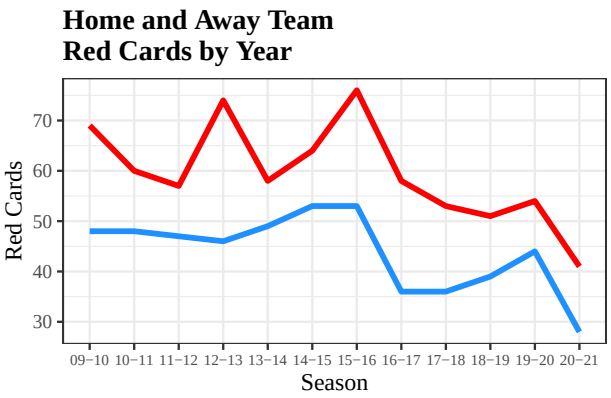

Figure 6

Figures 3 through 6 demonstrate that from 2009 to 2021 home teams have consistently gotten more shots and shots on goal (both of which are very widely used measures of team performance) than away teams. Additionally, away teams have consistently received more yellow cards than home teams, although the gap between the two has varied throughout the years. The increased number of yellow cards and red cards received by away teams can perhaps be attributed to the fact that referees are biased toward the home team and consequently may be more likely to call penalties on the away team. Another hypothesis is that away teams may be more likely to play aggressively due to being startled by the opposing team's stadium and crowd, and consequently commit more penalties.

2. Methodology

2.1 Linear Regression Model

From our exploratory data analysis, it was clear that in Serie A football, home teams have consistently performed better than away teams from 2009 to 2021. In order to further explore the reasons why home teams have performed better than away teams, we decided to create a linear regression model that predicts the score differential between the home and away teams. This model was fitted using the *lm()* function in R. By examining the intercept term for our model, we would then be able to see and quantify the home field advantage (or lack thereof) that persists when teams are evenly matched in terms of ranking, performance, and potential referee bias.

One thing to note is that when calculating the differential variables, we subtracted the statistic for the away team from the home team. Our initial "full" model will contain predictor variables corresponding to the following statistics:

- Difference in shots
- Difference in shots on target
- Difference in times a team hit the woodwork
- Difference in corner kicks
- Difference in yellow cards
- Difference in red cards
- Median-centered average percentage of home stadium filled
- Difference in team ranks at the time of the game
- Indicator variable for whether the game took place before (0) or after (1) the COVID-19 pandemic

After running our "full" model, we performed backwards selection using Akaike's Information Criterion (AIC) as our selection criterion. Additionally, we were concerned with high correlations between certain variables. For instance, we conjectured that there would be a high correlation between the variables for the difference in yellow cards between teams and the difference in red cards between teams. Curiously, there is not a concerning level of multicollinearity, as seen from the variance inflation factor (VIF) values, all of which are below 3. Since multicollinearity decreases statistical power, we were happily surprised to see that it is not a problem.

As seen in our exploratory data analysis, there are noticeable fluctuations in team statistics and behaviors that occur over time. For instance, yellow cards increase dramatically for the two seasons after both the 2010-2011 and 2017-2018 seasons. To control the effects of time on the predictors in this variable, we decided to include a random effect for the season. More specifically, we decided to include a random intercept. However, upon further examination, the addition of this random effect proved to be superfluous due to a low value for the Interclass Correlation Coefficient (ICC) (≈ 0.002).

Alternatively, we also explored including a categorical variable to account for the season, but again, this variable proved to be insignificant. Since the COVID-19 indicator variable was derived from the season variable, this variable had a high correlation with season. Our focus is the effect of the pandemic and not general seasonal differences, so we will use the variable for whether or not a game took place before or after the COVID-19 pandemic as the variable for adjusted fan attendance policies.

After deciding upon a final model, we examined diagnostics. Given that our final model is a linear regression model, the main assumptions of linear regression are as follows:

(1) Linearity: there is a linear relationship between the response and each predictor variable.

- This assumption is satisfied due to the fact that the residual plot for our model (see the Residuals vs Fitted plot in Figure 7 in the Appendix) does not have a discernible pattern or structure. We also see no discernible pattern in the standardized residuals of each continuous predictor variable (seen in Figures 8 through 15 in the Appendix).

(2) Constant Variance: the variability of the errors is equal for all values of the predictor variables.

- This assumption is satisfied because the vertical spread of the residuals is relatively constant across the residual plot (see the Residuals vs Fitted plot in Figure 7 in the Appendix).

(3) Normality: the errors follow a Normal distribution.

- This assumption is satisfied as the points fall along a straight diagonal line on the normal quantile (QQ) plot (see Figure 7 in the Appendix), indicating that the residuals follow a Normal distribution.

(4) Independence: the errors are independent from each other.

- Upon first glance, it would seem that the independence assumption is violated due to the fact that the data was collected over time. Additionally, game results in all sports, not just football, are dependent on each other. A thrilling victory can often give a team much-needed confidence and can be the catalyst for a lengthy winning streak. On the other hand, a crushing defeat can be detrimental to team morale and spiral into a series of losses. In the future, to control for potential team "momentum", we could look to include a way to control for winning and losing streaks. Additionally, one game can impact

another if a player (especially a star player) is injured. Unfortunately, we did not have access to player injury data, so injuries were things for which we were unable to control.

- Additionally, changes in personnel and general player ability can also cause games from the same season to be correlated with each other. However, as discussed earlier, any approach to control for the season in which a game is player, either through a random effect or a categorical variable, proved to be statistically insignificant.
- Since the data were collected in a particular order, we can examine a boxplot and scatterplot of the standardized residuals versus order in which the data were collected. In this case, information collected before COVID-19 were obviously collected before information collected during COVID-19. Looking at the boxplot (see Figure 16 in the Appendix), we can see that while there are a few more outliers in the standardized residuals for values before the pandemic, the quartiles are relatively similar. Furthermore, the scatterplot (see Figure 17 in the Appendix) shows no clear group patterns. As such, the independence condition is satisfied.

3. Results

3.1 Linear Regression Results

After attempting various model linear models both with and without mixed effects, we arrived at the following model:

$$\begin{aligned}\text{Goal Differential}_i =& 0.251 - 0.045(\text{Shot Differential})_i + 0.276(\text{Shots on Target Differential})_i \\ &+ 0.015(\text{Shots Hit Woodwork Differential})_i - 0.074(\text{Corner Kick Differential})_i \\ &- 0.042(\text{Yellow Card Differential})_i - 0.397(\text{Red Card Differential})_i \\ &+ 0.003(\text{Median-Centered Average Percentage of Home Stadium Filled})_i \\ &- 0.060(\text{Team Rank Differential})_i - 0.208(\text{Game Took Place After COVID-19})_i\end{aligned}$$

Table 1: Model Predictors

Predictor Variable	Coefficient Estimate	Standard Error	P-Value	95% Confidence Interval
(Intercept)	0.251	0.022	<0.001	0.207, 0.295
Shot Differential	-0.045	0.004	<0.001	-0.053, -0.037
Shots on Target Differential	0.276	0.008	<0.001	0.261, 0.291
Shots Hit Woodwork Differential	0.015	0.004	<0.001	0.008, 0.022
Corner Kick Differential	-0.074	0.005	<0.001	-0.085, -0.063
Yellow Card Differential	-0.042	0.012	<0.001	-0.066, -0.018
Red Card Differential	-0.397	0.041	<0.001	-0.477, -0.318
Median-Centered Average Percentage of Home Stadium Filled	0.003	0.001	0.003	0.001, 0.004
Team Rank Differential	-0.060	0.003	<0.001	-0.066, -0.055
Game Took Place After COVID-19	-0.208	0.072	0.004	-0.348, -0.067

All variables and the intercept are significant at the $\alpha = 0.05$ level.

3.2 Interpretations

Each of the explanatory variables (with the exception of the indicator variable for COVID-19, and the variable that measures home team stadium attendance) measures the difference in results between the home team and the away team. Furthermore, since our analysis is focused on the effect of the COVID-19 pandemic on

home-field advantage, we can state that when controlling for variables that correspond to fan support, referee bias, and coaching/team behavior, a home team playing during the pandemic is expected to make 0.208 less goals than the away team when compared to a home team playing prior to COVID-19. Looking at the intercept, one can see that if two teams are evenly matched in terms of offensive playing style, yellow and red cards, and team rank prior to the pandemic, we can expect the home team to win by about 0.251 goals, on average. This, in effect, implies that there is some evidence that home-field advantage–while relatively minor–decreased during the COVID-19 pandemic compared to prior years. More specifically, we are 95% confident that home team advantage–as measured by the difference in home and away team goals–decreased by 0.362 to 0.555 goals during COVID-19 when compared to the home advantage pre-COVID-19.

4. Discussion

4.1 Conclusions

In this case study, we sought to evaluate the statistically significant factors in Serie A home field advantage and the effect of the COVID-19 pandemic on this phenomenon given that stadiums severely limited the number of spectators allowed. Our model results provide several insights on the effects of fan support, referee bias, and coaching/team behavior and how their impacts on home field advantage have changed since the COVID-19 pandemic. Given that all the coefficients of the independent variables in our model are statistically significant at the $\alpha = 0.05$ level, we can conclude that there is sufficient evidence that these variables have non-zero correlations with score differential between the home and away teams.

Because our model intercept is interpreted as the home team advantage when all the differential variables are set at 0, we decided to center average percentage of home stadium filled at its median since it is unreasonable to infer that there are no spectators at a game pre-COVID. Our findings support our hypothesis that difference in number of shots, shots on targets, shots hit woodwork, corner kicks, number of yellow and red cards, and team rank, and the median-centered average percentage of home stadium filled s are associated with the home team winning.

The intercept of our model indicates that the expected goal differential between home and away teams is 0.251 and that we are 95% confident that the average goal differential between home and away teams is between 0.207 and 0.295 when all the differential variables are set to 0, the average percentage of home stadium filled is the median value, and the game takes placed before the COVID-19 pandemic. Therefore, we can conclude that home field advantage does indeed exist.

Holding all other variables constant, our model also suggests that the COVID-19 pandemic has a negative correlation with score differential, as home teams are expected to score 0.208 less goals during games that took place during the pandemic (or in other words, after the start of the pandemic in 2020). This finding is not surprising due to the fact that there were significantly less (and in some cases, no fans) at games after the start of the pandemic.

4.2 Limitations and Future Directions

Our approach has a few limitations stemming from the difficulty in quantifying home field advantage given that player, coach, and referee behavior may be influenced by psychological or other factors that we cannot necessarily quantify. Additionally, our model only utilizes data from 12 seasons of one league. If we were to incorporate data from other leagues from all over the world, we could better identify the extent of certain factors contributing to home field advantage and trends that may be more or less prevalent in certain leagues compared to others. Another key assumption in our model is that when interpreting our explanatory variable for the average percentage of home stadium filled, the variable assumes that the spectators are fans of the home team and are therefore contributing to the "12th man" phenomenon. In other words, we did not consider whether or not the spectators root for the away team and perhaps detract from the home team's "advantage". Lastly, there are a few observations, or matches, that are noticeably more influential than the others, as seen in Figure 18 of the Appendix. These matches are likely from 2020 and are likely more influential due to the fact that they are the first matches that take place after the COVID-19 pandemic.

Nonetheless, these points are not concerning as they are not *too* influential (<0.01).

If we were to approach the study again, one additional aspect we could investigate is the distance that away teams have to travel to play a match. While the distance that away teams usually have to travel is not as far as in Serie A than in other leagues, distance traveled could potentially be a variable contributing to home field advantage. Additionally, teams may sometimes play multiple away matches in a row, which can impact how well a team maintains its practice and rest schedule or familiarity with its environment. However, because the datasets we used were limited to the scope of what happens during a match, it could be interesting to see if the distance traveled by the away team is a statistically significant variable in goal differential between home and away teams.

Another variable we would include in our model is injury time. As mentioned previously, prior studies have assessed referee favoritism towards the home team by looking at injury times as a measure of social pressures in referee behavior. It could be useful to include injury time as an explanatory variable to further control for potential referee bias.

Linear regression is not the only method for modeling this data. Alternative methods of modeling this data include Poisson regression and Negative Binomial regression. Poisson regression would be useful if modeling a non-negative integer such as the number of goals or shots by the home team since it is used with count data. Negative binomial regression would also be useful if overdispersion exists in said variable. Both types of regression are also highly interpretable and would allow us to determine the magnitude of home field advance based on the statistical significance and magnitude of an intercept term.

4.3 Summary

Overall, our case study demonstrates that home team advantage–while relatively small–is observed in Serie A seasons from 2009 until the pandemic. Since the pandemic when spectators were not allowed in stadiums, home team advantage is negligible, supporting previous studies on the impact of the 12th man phenomenon on coaching, referee, and player behavior. Given the instability of the COVID-19 pandemic and the uncertainty it creates in policies on travel, spectators allowed, and other factors, these findings can help teams and sports betters to have a better understanding on how to adjust their strategies during these times.

References

Bilalić, Merim, et al. "Home Advantage Mediated (HAM) by Referee Bias and Team Performance during Covid." Scientific Reports, vol. 11, no. 1, Dec. 2021, p. 21558. DOI.org (Crossref), https://doi.org/10.1038/s41598-021-00784-8.

Garicano, Luis, et al. "Favoritism under Social Pressure." The Review of Economics and Statistics, vol. 87, no. 2, 2005, pp. 208–16. JSTOR, https://www.jstor.org/stable/40042898.

"Generalised Linear Models with Glm and Lme4." Rens van de Schoot, https://www.rensvandeschoot.com/tutorials/generalised-linear-models-with-glm-and-lme4/. Accessed 22 Jan. 2022.

Goumas, Chris. "Home Advantage and Referee Bias in European Football." European Journal of Sport Science, vol. 14, no. sup1, Jan. 2014, pp. S243–49. DOI.org (Crossref), https://doi.org/10.1080/17461391.2012.686062.

Hallé Petiot, Grégory, et al. "Contrasting Learning Psychology Theories Applied to the Teaching-Learning-Training Process of Tactics in football." Frontiers in Psychology, vol. 12, May 2021, p. 637085. PubMed Central, https://doi.org/10.3389/fpsyg.2021.637085.

Jamieson, J. P. (2010). The home field advantage in athletics: a meta-analysis. J. Appl. Soc. Psychol. 40, 1819–1848. doi: 10.1111/j.1559-1816.2010.00641.x

Marek, Patrice, and František Vávra. "Comparison of Home Advantage in European Football Leagues." Risks, vol. 8, no. 3, Aug. 2020, p. 87. DOI.org (Crossref), https://doi.org/10.3390/risks8030087.

Wunderlich, Fabian, et al. "How Does Spectator Presence Affect Football? Home Advantage Remains in European Top-Class Football Matches Played without Spectators during the COVID-19 Pandemic." PLOS ONE, vol. 16, no. 3, Mar. 2021, p. e0248590. PLoS Journals, https://doi.org/10.1371/journal.pone.0248590.

Appendix

Table 2: Linear Regression Output

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	0.251	0.207, 0.295	**<0.001**
Shots.Diff	-0.045	-0.053, -0.037	**<0.001**
Shots.on.Target.Diff	0.276	0.261, 0.291	**<0.001**
Woodwork.Diff	0.015	0.008, 0.022	**<0.001**
Corner.Diff	-0.074	-0.085, -0.063	**<0.001**
Yellow.Diff	-0.042	-0.066, -0.018	**<0.001**
Red.Diff	-0.397	-0.477, -0.318	**<0.001**
Capacity.Percent.Centered	0.003	0.001, 0.004	**0.003**
CurrentResult.Diff	-0.060	-0.066, -0.055	**<0.001**
as.factor(Covid)			
0	—	—	
1	-0.208	-0.348, -0.067	**0.004**

[1]CI = Confidence Interval

Table 3: Linear Mixed Model Output

Characteristic	Beta	95% CI[1]
(Intercept)	0.251	0.202, 0.301
Shots.Diff	-0.045	-0.053, -0.037
Shots.on.Target.Diff	0.276	0.261, 0.291
Woodwork.Diff	0.015	0.008, 0.022
Corner.Diff	-0.074	-0.085, -0.064
Yellow.Diff	-0.042	-0.066, -0.018
Red.Diff	-0.397	-0.477, -0.317
as.factor(Covid)		
0	—	—
1	-0.208	-0.371, -0.045
Capacity.Percent.Centered	0.003	0.001, 0.005
CurrentResult.Diff	-0.061	-0.066, -0.055
Season.sd___(Intercept)	0.040	

[1]CI = Confidence Interval

Characteristic	Beta	95% CI[1]
Residual.sd___Observation	1.25	

[1]CI = Confidence Interval

Table 4: Interclass Correlation Coefficients

Metric	Value
Adjusted ICC	0.002
Conditional ICC	0.001

Table 5: Variance Inflation Factors

Variable	VIF
Difference in Shots	2.798
Difference in Shots on Target	2.215
Difference in Hit Woodwork	1.102
Difference in Corners	1.548
Difference in Yellow Cards	1.126
Difference in Red Cards,	1.048
Median-Centered Average Percentage of Home Stadium Filled	1.047
Difference in Team Ranks	1.352
Before (0) or After (1) Covid	1.038

Diagnostic Plots

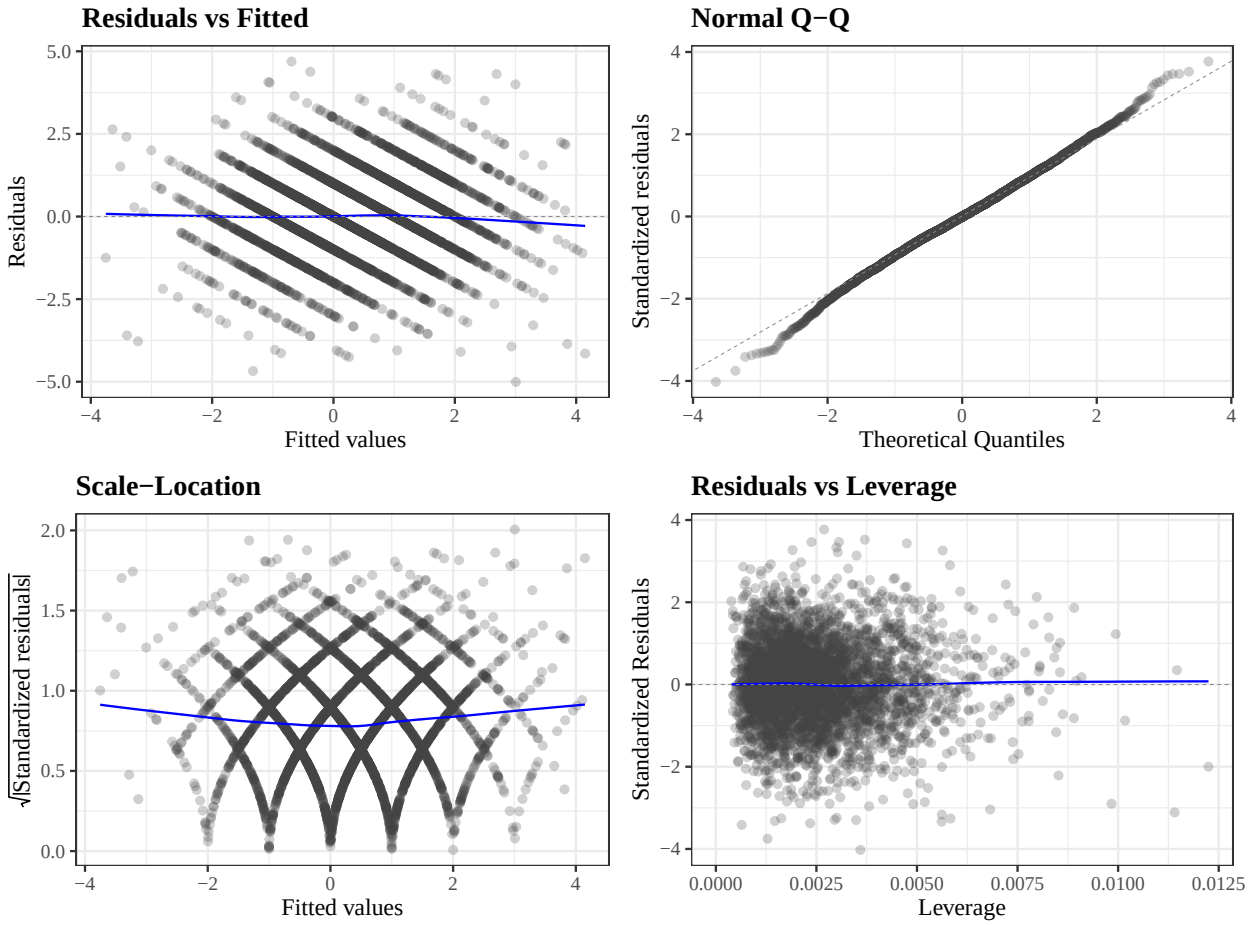

Figure 7

Standardized Residuals vs Each Predictor

Standardized Residual vs Shot Differential

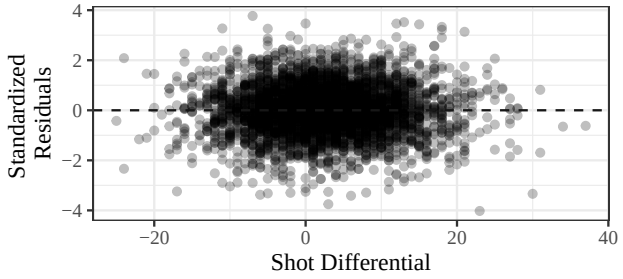

Figure 8

Standardized Residual vs Shots on Target Differential

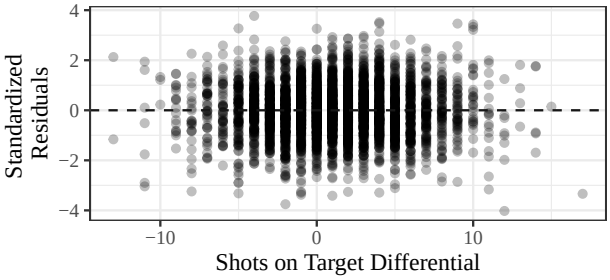

Figure 9

Standardized Residual vs Shots Hit Woodwork Differential

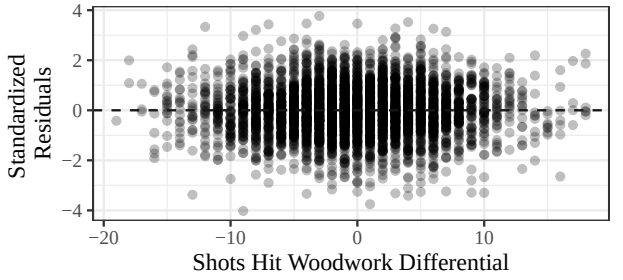

Figure 10

Standardized Residual vs Corner Kick Differential

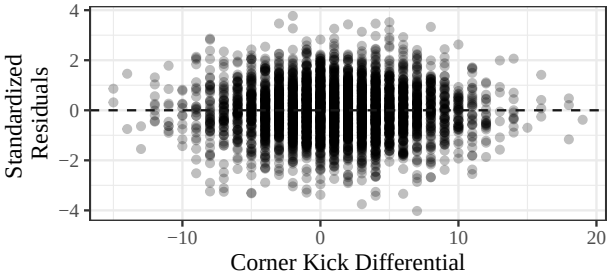

Figure 11

Standardized Residual vs Yellow Card Differential

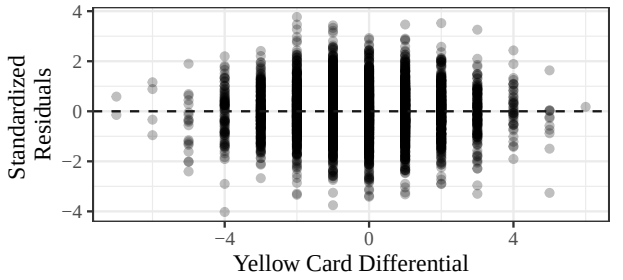

Figure 12

Standardized Residual vs Red Card Differential

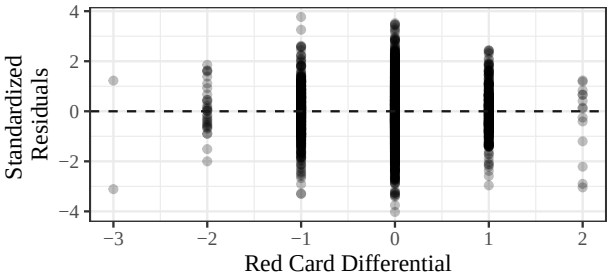

Figure 13

Standardized Residual vs Median−Centered Average Percentage of Home Stadium Filled

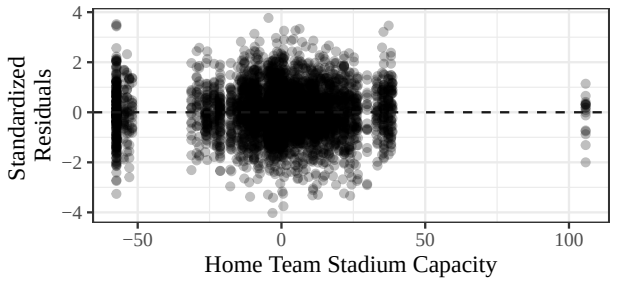

Figure 14

Standardized Residual vs Team Rank Differential

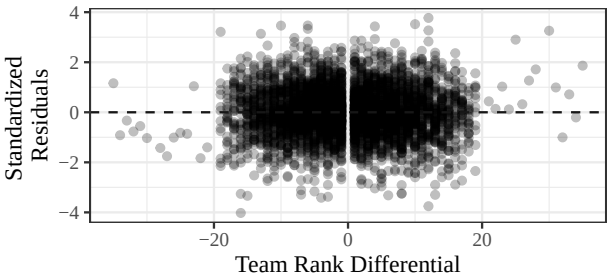

Figure 15

Standardized Residuals by COVID Presence

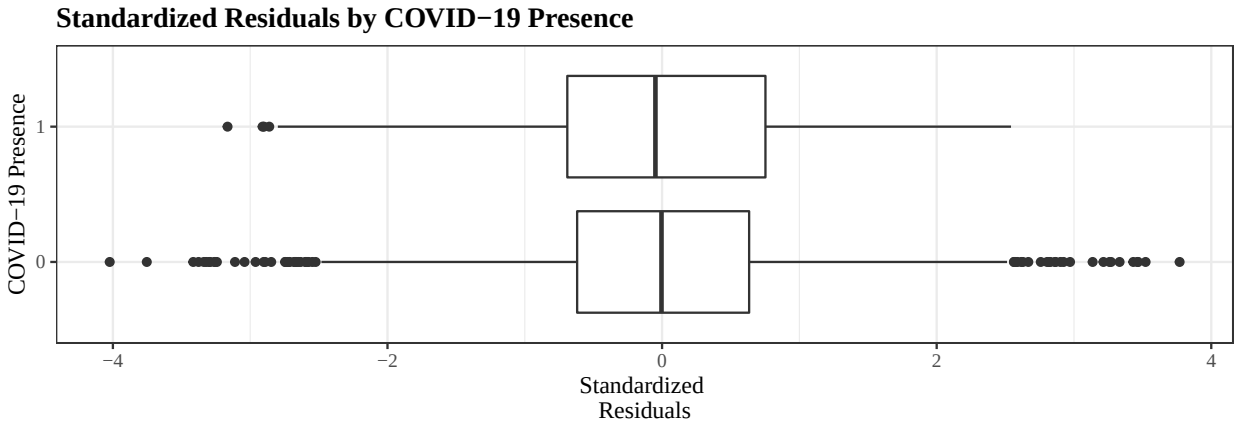

Figure 16

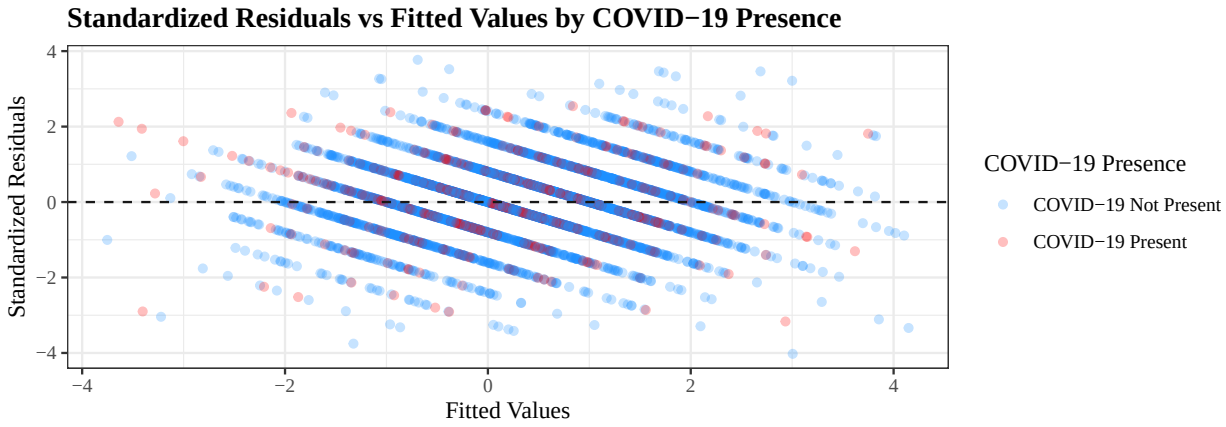

Figure 17

Leverage, Cook's Distance, and Standardized Residuals

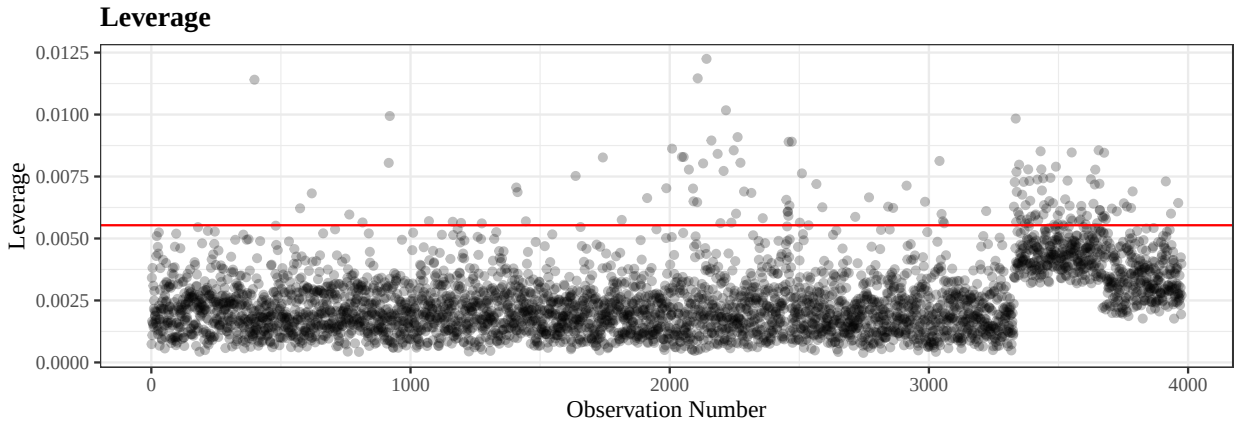

Figure 18

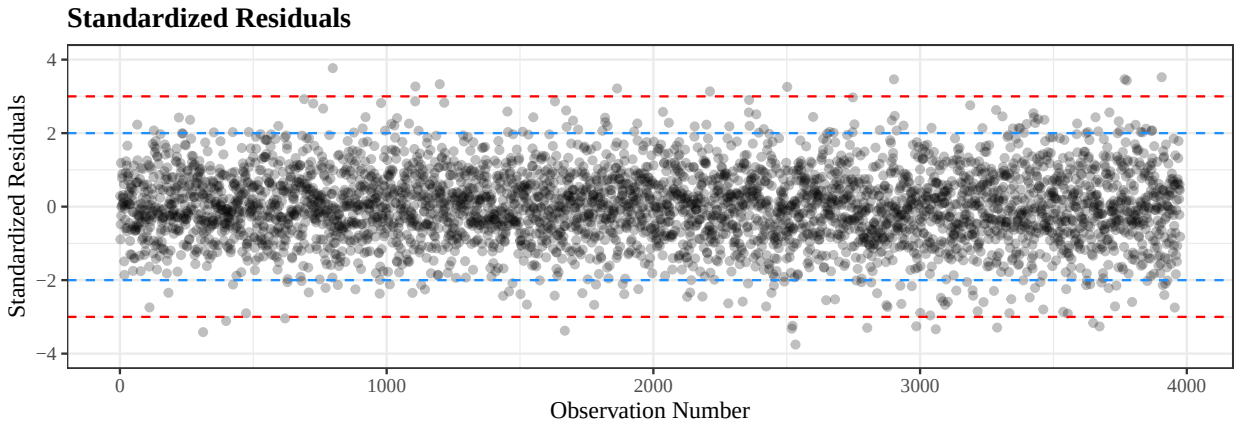

Figure 19

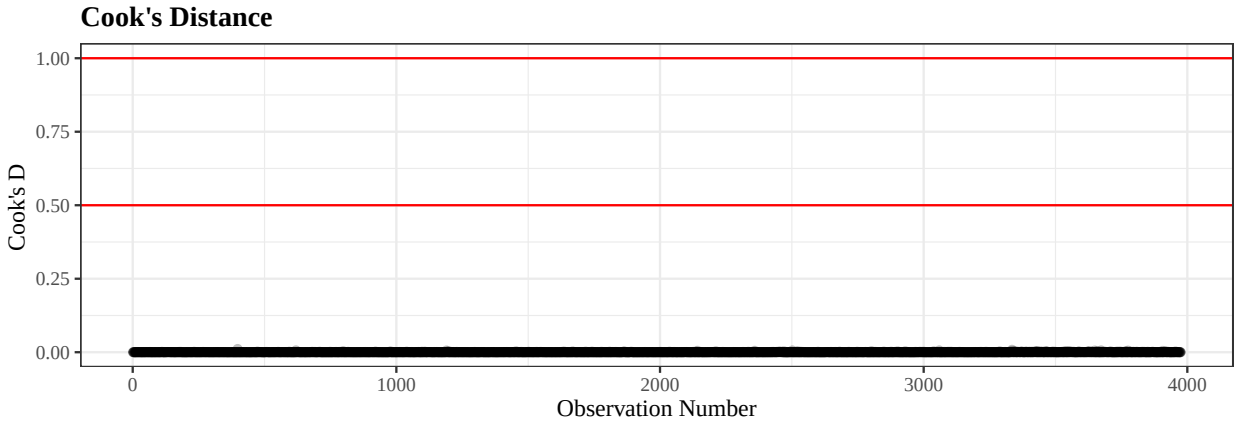

Figure 20