Estimating the Probability that FC Barcelona Win a Game Using a Logistic Model

Abbass Sleiman

Table of contents

Introduction

In this paper, we seek to diverge from Maher's 1982 paper (Maher 1982), and instead implement a logistic regression in order to estimate the probability that FC Barcelona, winner of the 2022-2023 LaLiga, win a game given that a game is Home or Away, the amount of possession they held, who the captain was between one of Sergio Busquets, Marc-André ter Stegen, Sergi Roberto, and Gerard Piqué, as well as what day of the week the game was played.

Data

The data that this paper utilizes contains information all 38 games played by FC Barcelona in the 2022-2023 LaLiga season, including the result of the game (win, draw, or loss), what day of the week the game was played, the percentage of possession that FC Barcelona held throughout the game (between 0 and 100), as well as which of the aforementioned captains was in charge of each particular game. The data was collected from Sports Reference's logs and then manually converted into an excel file which was then turned into a csv file for the purposes of this paper (Reference 2023).

Minimal data cleaning was required except for mutating the data such that a win registers a value of "1", and a draw or loss register a value of "0" for use in the regression.

Note that all the data analysis was done through R (R Core Team 2023) with the aid of the following packages: tidyverse (Wickham et al. 2019), here (J. B. Kirill Müller 2020), dplyr (Hadley Wickham 2023), tibble (R. F. Kirill Müller Hadley Wickham 2023), janitor (Sam Firke 2023), knitr (Xie 2023), kableExtra (Zhu 2021), broom (Alexander, Collingwood, and Whitford 2022), rstanarm (Gabry et al. 2022), and marginaleffects (Solymos 2021).

Model

The particular model that this paper will utilize is logistic. The primary reason why I have decided to make use of a logistic model in particular, is due to the binary nature of the outcome variable that we are interested in - whether FC Barcelona will win or lose. Though there is the possibility of a draw, as this paper is primarily concerned with whether or not FC Barcelona wins, we treat a draw as a loss in this regard.

The independent variables which we are interested in examining with regards to their effect on the probability that FC Barcelona wins or loses a game are as follows: Day (one of Saturday, Sunday, Monday, Tuesday, Wednesday, or Thursday), Venue (Home or Away), Possession (integer value between 0-100), and one of the aforementioned captains.

Thus, the model is as follows:

$$\begin{aligned} y_i|\pi_i &\sim \text{Bern}(\pi_i) \\ \text{logit}(\pi_i) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} \end{aligned}$$

Here,

$$y_i = 1$$

denotes a win, and each of the independent variables from 1 to 4 represent the day, venue, possession, and the captain respectively.

Results and Discussion

After running our regression, we get the values for the coefficients of each of our independent variables as showcased in Table 1. We also compare the predictions made by the model for the likelihood that FC Barcelona wins a game, compared with the actual result for the first 6 games in Table 2. It is clear that as it stands, the model is quite ineffective in its predictive power, judging by the incredibly large standard errors and p-values present in Table 1. This is likely due to the small number of observations (only 38), and even smaller number of observations when categorizing each observation by day, captain, and possession.

Table 1: Coefficients for each of the independent variables calculated from running the logistic regression.

term	estimate	std.error	statistic	p.value
(Intercept)	3.1968517	5594.8854224	0.0005714	0.9995441
daySat	20.3470108	3956.1805468	0.0051431	0.9958964
daySun	20.1879803	3956.1804770	0.0051029	0.9959285
dayThu	37.8052070	5594.8840325	0.0067571	0.9946087
dayTue	20.4525671	3956.1807904	0.0051698	0.9958751
dayWed	18.9035594	3956.1808118	0.0047782	0.9961875
venueHome	1.0094036	0.9863152	1.0234088	0.3061146
poss	-0.0998171	0.0712843	-1.4002681	0.1614331
captainMarc-André ter Stegen	-16.0402034	3956.1806537	-0.0040545	0.9967650
captainSergi Roberto	-17.5580133	3956.1805414	-0.0044381	0.9964589
captainSergio Busquets	-15.8831144	3956.1804715	-0.0040148	0.9967967

Table 2: Predicted likelihood of FC Barcelona winning a game compared with the actual result.

Game	Estimated Probability	Result	Day	Venue	Possession	Captain
1	0.8789373	0	Sat	Home	67	Sergio Busquets
2	0.8256687	1	Sun	Away	58	Marc-André ter Stegen
3	0.8609706	1	Sun	Home	67	Sergio Busquets
4	0.9064148	1	Sat	Away	54	Sergio Busquets
5	0.6622977	1	Sat	Away	70	Sergio Busquets
6	0.7164586	1	Sat	Home	76	Marc-André ter Stegen

References

Alexander, Leigh, Loren Collingwood, and Andrew B. Whitford. 2022. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Gabry, Jonah, Ben Goodrich, Imad Ali, Jeffrey Arnold, Matthew Kay, Michael Betancourt, and Aki Vehtari. 2022. *rstanarm: Bayesian Applied Regression Modeling via Stan*. https://mc-stan.org/rstanarm/.

Hadley Wickham, Lionel Henry, Romain François. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Kirill Müller, Jennifer Bryan. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Kirill Müller, Romain Francois, Hadley Wickham. 2023. *Tibble: Simple Data Frames*. https://CRAN.R-project.org/package=tibble.

Maher, Michael. 1982. "Modelling Association Football Scores." *Statistica Neerlandica* 36 (3): 109–18. https://doi.org/10.1111/j.1467-9574.1982.tb00782.x.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Reference, Sports. 2023. *2022-2023 Barcelona Scores and Fixtures (La Liga)*. https://fbref.com/en/squads/206d90db/2022-2023/matchlogs/c12/schedule/Barcelona-Scores-and-Fixtures-La-Liga.

Sam Firke, Chris Haid, Bill Denney. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Solymos, Peter. 2021. *Marginaleffects: Marginal Effects for Model Objects*. https://CRAN.R-project.org/package=marginaleffects.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the Tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.