Simplified Model of Maher's "Modelling association football scores"*

Using Poisson Regression to Model Manchester United's Home and Away Games in the 22/23 season

Krishiv Jain

March 18, 2024

1 Introduction

This paper aims to build a simplified version of Maher's 1982 "Modelling association football scores" paper (Maher (1982)). In this paper, Maher uses the Poisson Regression to model the number of goals scored in a football (soccer) games. This paper uses his paper and model as a guide to conduct a similar analysis. To conduct his analysis, Maher looks at each team's attacking strength and defensive weakness, both home and away. Maher considered a single game played as giving 2 observations: one for the number of goals scored by the home team, and one for the number of goals scored by the away team. Through a Poisson Regression, and using the aforementioned four metrics, Maher estimated the scores of 462 games (the league had 22 teams), giving him 924 data points. Maher's paper was motivated by attempting to justify the Poisson Regression as being a valid model for modelling football scores. Thus, he delves into various intricacies, such as having four different models whereby each model has a different set of constraints on the predictors.

In this paper, the focus is placed on the 2022-2023 Premier League season. The same concept as Maher for the predictors are used, except a metric for defensive strength, instead of weakness, is used. The formulas to calculate these four metrics were taken from SBO.net ("Poisson Distribution" (n.d.)). Moreover, this paper focuses on Manchester United, and using a poisson regression to estimate the number of goals they would have scored against a team. 2 models are used; one for United playing at home, and one for them playing away.

*Code and data are available at: https://github.com/Krishiv-J/Tutorial-10

2 Data

2.1 Raw Data

2 sets of raw data were created. The first one was from whoscored.com (*Premier League Team Statistics* (2024)). Data was provided for various leagues, and for different seasons. For the purpose of this project, the most recent season (2022-2023) for the Premier League was chosen. This source provides information on a variety of metrics for each team. The data that was needed for this paper was the total number of goals scored and the number of goals conceded at home, and away from home for each team. Thus, there was a total of 20 rows (since there are 20 teams) and for each team, there were 4 variables.

The second data set created was from worldfootball.net (*Manchester United » Fixtures & Results 2022/2023* (2024)). This website provided the results of every fixture played during the 2022-2023 season. For the purpose of this paper, only the premier league games were needed. Moreover, only the number of goals scored in each game by United were needed.

2.2 Clean Data

After reading the data from the first data set, 2 data tables were created: one for home games, one for away games. Each data table included the number of goals scored and conceded. Next, the average number of goals scored and conceded per game was calculated for both home, and away games. Then, the league's average for each of the four metric was calculated. Finally, a team's home attacking strength was calculated as the average number of goals they scored at home divided by the league's average. This same concept was used to calculate each team's home attacking strength, home defensive strength, away attacking strength, away defensive strength. The row for Manchester United was removed since the information was not needed for our model.

The second raw data table was read in. This included information on how many goals United scored against each team in their home and away games. This was merged with the clean data from before.

3 Model

The goal of our modelling strategy is twofold. Firstly,…

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in Appendix B.

Table 1: Explanatory models of flight time based on wing width and wing length

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \alpha + \beta_i + \gamma_i \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\gamma \sim \text{Normal}(0, 2.5) \tag{5}$$
$$\sigma \sim \text{Exponential}(1) \tag{6}$$

We run the model in R (R Core Team 2024) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular...

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in Table 1.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

Examining how the model fits, and is affected
by, the data

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

Checking the convergence of the MCMC algo-
rithm

References

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

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

Manchester United » Fixtures & Results 2022/2023. 2024. worldfootball.net. https://www.worldfootball.net/teams/manchester-united/2023/3/.

"Poisson Distribution." n.d. https://www.sbo.net/strategy/football-prediction-model-poisson-distribution/.

Premier League Team Statistics. 2024. whoscored.com. https://www.whoscored.com/Regions/252/Tournaments/2/Seasons/9075/Stages/20934/TeamStatistics/England-Premier-League-2022-2023.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.