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 19, 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.

The data was downloaded, cleaned, and analyzed using the statistical programming language R (R Core Team 2024). The following packages were also used: janitor (Sam Firke 2023),

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

here (Kirill Müller 2020), as well as tidyverse (Wickham et al. 2019), which includes the dplyr (Hadley Wickham 2023).

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 Poisson regression is used to stay consistent with Maher. Maher's reasoning for using this mode was that the Poisson regression is an appropriate model given that the number of goals

Table 1: Regression models of home goals scored based opposition's away attacking and defending strength

	Home model
(Intercept)	0.153
	(1.112)
away_attacking_strength	0.000
	(0.756)
away_defending_strength	0.358
	(0.528)
Num.Obs.	19
AIC	63.2
BIC	66.0
Log.Lik.	−28.591
RMSE	0.99

scored in a football game is considered count data. Moreover, a Poisson regression aligns with the nature of the data. This distribution is particularly applicable due to the reasoning that the number of goals scored by a team in a match is likely to follow a Poisson variable. In football, possession plays a crucial role, offering teams repeated opportunities to attack and score. Whilst the probability of scoring during each possession may be small, the large number of possession opportunities throughout a match leads to a scenario where the number of goals follows a Poisson distribution. (Maher (1982))

In this paper, two separate Poisson regressions are conducted. One is for the matches that United played at home, and one for the games that they played away. For both models, the outcome variable is the number of goals scored by United. For the games played at home, the independent variables are the opposition team's `away_attacking_strength` and `away_defending_strength`. For the games played away, the independent variables are the opposition team's `home_attacking_strength` and `home_defending_strength`.

4 Results

Table 1 shows the results from the first model, whilst Table 2 shows the results from the first model. To further explore the results, the values for the estimates of the coefficients were used to estimate the number of goals scored in each game that United played. Table 3 showcases this for the home games, alongside the actual goals that were scored, whilst Table 4 showcases this for the away games, alongside the actual goals that were scored.

Table 2: Regression models of away goals scored based opposition's home attacking and defending strength

	Away model
(Intercept)	−2.310
	(1.399)
home_attacking_strength	0.844
	(0.602)
home_defending_strength	1.500
	(0.863)
Num.Obs.	19
AIC	52.1
BIC	54.9
Log.Lik.	−23.049
RMSE	0.74

Table 3: Predicted and Actual Goals scored at home against each team

team	predicted_home_goals	actual_home_goals
Arsenal	1.53813063258533	3
Aston Villa	1.71426954820989	1
Bournemouth	2.26451264052583	3
Brentford	1.795586474628	1
Brighton	1.90994051800821	1
Chelsea	1.79567473453216	4
Crystal Palace	1.74096896116198	2
Everton	1.8520812985924	2
Fulham	1.68786344410769	2
Leeds	2.19543724783713	2
Leicester	2.19530775793012	3
Liverpool	1.85188103329548	2
Manchester City	1.4913003383252	2
Newcastle	1.56214814207468	0
Nottingham Forest	2.29994024382823	3
Southampton	2.03217656572223	0
Tottenham	2.09568258224306	2
West Ham	1.88098257489174	1
Wolves	2.09611526150063	2

Table 4: Predicted and Actual Goals scored away against each team

team	predicted_away_goals	actual_away_goals
Arsenal	2.11838651330272	2
Aston Villa	0.949075254946886	1
Bournemouth	1.0491424374023	1
Brentford	0.825046049899412	0
Brighton	1.05809723745402	0
Chelsea	0.585503743729678	1
Crystal Palace	0.779678951946965	1
Everton	0.881991802822275	2
Fulham	1.50955168727599	2
Leeds	2.21297561233932	2
Leicester	1.06686182868369	1
Liverpool	1.0428018000255	0
Manchester City	1.52576984724761	3
Newcastle	0.654190932934051	0
Nottingham Forest	0.979260373820914	2
Southampton	1.82950246925329	1
Tottenham	1.37121476777737	2
West Ham	0.952997910435216	0
Wolves	0.607950778740702	1

References

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.

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/.

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

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.