Modelling Football Scores for the 22/23 Season of the English Premier League*

My subtitle if needed

Jiwon Choi

March 19, 2024

1 Introduction

Predicting the outcomes of sports events, particularly football (soccer) matches, has long been a subject of interest not only among fans and betting enthusiasts but also within the academic community. The inherent uncertainty and variability in sports outcomes, influenced by numerous factors ranging from team strength and strategy to individual player performance and even unforeseen events, make accurate prediction a challenging yet fascinating problem.

In association football, the scoring dynamics and match results can be significantly influenced by both observable factors such as team composition and historical performance, and less quantifiable elements like team morale or even weather conditions. Traditional approaches to predicting football match scores have often relied on Poisson regression models due to the Poisson-like distribution of goals scored in matches. These models, while useful, assume a fixed rate of goal-scoring that might not fully capture the complexities of football matches.

The advent of Bayesian statistical methods offers a new perspective, allowing for the incorporation of prior knowledge and the estimation of model parameters with a degree of uncertainty. This approach acknowledges the inherent unpredictability in sports outcomes and provides a probabilistic framework for prediction.

This report introduces a Bayesian Poisson model for predicting the number of goals scored by home teams in the English Premier League for the 2022-2023 season. Developed using the `rstanarm` package (Goodrich et al. (2022)), this model leverages prior distributions and Bayesian inference to estimate goal-scoring rates, offering a more nuanced understanding of match outcomes. The objective is not only to predict the number of goals scored in a match but

*Code and data are available at: https://github.com/jwonc4602/Modelling-Football-Scores.

also to explore the factors contributing to home team advantage and to evaluate the predictive performance of Bayesian models in the context of football matches.

By embracing a Bayesian approach, this analysis aims to contribute to the broader discourse on sports prediction methodologies, offering insights that could benefit both academic research and practical applications in sports analytics and betting strategies.

2 Data

The dataset analyzed in this report consists of match results and statistics from the English Premier League for the 2022-2023 season. The data were sourced from the comprehensive football data repository at Football-Data.co.uk, a widely recognized resource for football match data across major European leagues and competitions.

The dataset includes detailed match data covering a wide array of variables, from basic match outcomes to more intricate betting odds provided by multiple bookmakers. Specifically, it encompasses match dates, team names, full-time and half-time scores, and various statistics that reflect team performance, such as shots on target and possession percentages. Additionally, the dataset features betting odds from several bookmakers, offering insights into market expectations regarding match outcomes.

A detailed description of the variables included in the dataset is provided in the accompanying `note.txt` file, which serves as a dictionary explaining each variable's significance and format. Key variables of interest for our analysis include:

- `Date`: The date on which the match was played.
- `HomeTeam` and `AwayTeam`: Names of the teams playing, with one designated as the home team and the other as the away team.
- `FTHG` (Full Time Home Goals) and `FTAG` (Full Time Away Goals): The number of goals scored by the home and away teams by the end of the match.
- `FTR` (Full Time Result): The outcome of the match, indicated as H (home win), D (draw), or A (away win).

Prior to analysis, the dataset underwent a cleaning process to ensure data quality and relevance to the study's objectives. This process involved selecting the variables of interest, handling missing values, and converting data into appropriate formats for analysis. Specifically, missing values in categorical variables such as `HomeTeam` and `AwayTeam` were marked as "N/A" to maintain the integrity of the dataset without discarding valuable match records. The cleaned dataset was saved as `cleaned_data.csv` for subsequent analysis.

The cleaned dataset forms the foundation for developing a Bayesian Poisson model to predict home team goal-scoring rates in the English Premier League, leveraging the rich match data to uncover patterns and insights into the factors influencing match outcomes.

3 Model

In exploring association football scores, we find compelling reasons to model the number of goals scored by a team as a Poisson variable. This stems from the nature of football, where possession is crucial, and each possession represents an opportunity to attack and score. Despite numerous attacks in a match, the probability of any single attack resulting in a goal is low. If the probability of scoring (p) remains constant and attacks are independent, the number of goals aligns with a Binomial distribution, which approximates a Poisson distribution for a large number of attacks with a small success probability.

Given the variability in team quality, we adopt an independent Poisson model. When team i plays at home against team j, with the observed score being (x_{ij}, y_{ij}), we assume X_{ij} (home team goals) and Y_{ij} (away team goals) are independent Poisson variables with means λ_{ij} and μ_{ij} respectively. The parameters λ and μ represent the attacking strength of the home team and the defensive strength of the away team, modulated by their respective abilities.

We define $\lambda_{ij} = \alpha_i \beta_j$ and $\mu_{ij} = \gamma_i \delta_j$, where:

- α_i represents the attacking strength of team i when playing at home.
- β_j represents the defensive weakness of team j when playing away.
- γ_i represents the defensive weakness of team i when playing at home.
- δ_j represents the attacking strength of team j when playing away.

Under these definitions, the likelihood function for the observed scores across all matches is:

The maximum likelihood estimates (MLEs) for the parameters can be obtained through iterative methods, like the Newton-Raphson technique. Initial estimates can be significantly improved by alternating between estimating α's with fixed β's and vice versa. This iterative process can be mathematically represented as follows:

$$L(\alpha, \beta, \gamma, \delta) = \sum_{i,j} \left[x_{ij} \log(\lambda_{ij}) - \lambda_{ij} + y_{ij} \log(\mu_{ij}) - \mu_{ij} \right],$$

where the constraints $\sum \alpha_i = \sum \beta_i$ and $\sum \gamma_i = \sum \delta_i$ ensure a unique set of parameters.

To assess the necessity of all parameters for an adequate score description, we explore a hierarchy of models, ranging from all teams being identical in all aspects to a comprehensive model with independent parameters for different teams' offensive and defensive capabilities at home and away. The hierarchy and the model selection process based on log likelihood and goodness-of-fit tests are depicted in Table 1.

Table 1: Coefficients from the Model

term	estimate	std.error	conf.low	conf.high
(Intercept)	0.4905249	0.0405845	0.4249691	0.557768

4 Results

The Bayesian Poisson model was successfully developed to predict the number of goals scored by home teams in the English Premier League for the 2022-2023 season. The model was trained using `rstanarm`, allowing for a Bayesian approach to regression with Poisson-distributed outcomes. Summary statistics of the model indicate that the intercept, representing the log rate of goals scored by home teams, was estimated with a credible interval that excludes zero, suggesting a positive effect of being the home team on the number of goals scored. Predictions from the model were generated, providing a probabilistic view of home team performance across the season.

5 Discussion

The use of a Bayesian Poisson model represents a methodological advancement in predicting football match outcomes, leveraging prior knowledge and the probabilistic nature of Bayesian inference. The model's results underscore the home advantage in football, consistent with existing literature. However, this analysis was limited by the simplicity of the model, which only included an intercept as a predictor. Future work could enhance the model's predictive power and insights by incorporating more predictors, such as team strength, historical performance, and player-level data.

The model's performance and the accuracy of its predictions could be further evaluated against actual match outcomes or compared to traditional betting odds. Such an evaluation would provide a more concrete assessment of the model's utility in forecasting football scores and could inform strategies for sports betting or team management.

References

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