Exploring the Dynamics of Football Scoring: A Statistical Analysis Inspired by Maher's Model*

Junwei Chen

March 19, 2024

This study applies a simplified version of Maher's 1982 Poisson distribution model to recent English Premier League data to analyze football scores. Using the engsoccerdata dataset, we calculated team-specific attacking and defensive strengths and employed these in Poisson regression models to predict the number of goals scored by home and away teams. Our results validate Maher's approach, showing a significant correlation between team strengths and goal-scoring. However, the positive association between defensive strength and goals scored indicates complexities not fully addressed by the model. This analysis highlights the utility of statistical methods in sports analytics and suggests avenues for further research, including the development of more sophisticated models to better capture the dynamics of football scoring.

Table of contents

[*] Code and reporduction data are available at: https://github.com/JunweiChen1012/football_scoring

1 Introduction

Football is more than just a game; it's a global phenomenon that brings people together, sparking joy, excitement, and sometimes heartbreak. While the thrill of the match captures our attention, there's a whole world of numbers and patterns behind each game that can tell us more about what's happening on the pitch. Back in 1982, a researcher named Maher (Maher 1982) took a deep dive into these numbers. He tried to figure out if you could predict how many goals would be scored in a football match by using maths, specifically something called the Poisson distribution. This was a big deal because it opened up new ways to think about predicting football scores by looking at the strengths of the teams playing.

In this essay, we're going to take a closer look at how we can use some simple statistics to understand football scores better. We'll start with the basics of Maher's idea and see if we can apply it to football matches today. It's not just about guessing who will win or lose, but about seeing how the numbers can tell us about the chances of different scores happening. We'll also talk about why some teams seem to do better at home and how all of this can help us get a better grasp of the beautiful game.

By exploring these ideas, we hope to show that football is not only about passion and skill but also about the fascinating patterns that emerge when we pay attention to the scores and statistics. So, whether you're a die-hard fan, a curious observer, or somewhere in between, join us as we delve into the numbers behind the game and discover what they can teach us about football.

2 Data

To explore the statistical underpinnings of football scoring and apply Maher's model to contemporary football matches, we utilized a dataset from the `engsoccerdata` package (Curley 2020), which compiles historical match outcomes from English football leagues. This dataset includes a wealth of information spanning several decades, providing a robust foundation for our analysis. Key variables extracted for this study include:

- **Date**: The date on which each match was played, allowing us to track changes over time and identify any temporal patterns in football scoring.
- **Season**: The football season during which the match occurred, facilitating analysis of trends and performance across different seasons.
- **Home and Visitor**: The names of the home and away teams, respectively, which are crucial for assessing home advantage and comparing team strengths.
- **FT (Full Time Score)**: The full-time score of the match, split into `hgoal` (home goals) and `vgoal` (visitor goals), serving as our primary outcome variables for modeling football scores.

- **Division**: The league division in which the match was played, enabling us to segment our analysis by the level of competition.
- **Tier**: The tier of English football to which the division belongs, further categorizing matches by the competitive level.
- **Totgoal** and **Goaldif**: Aggregate metrics representing the total goals scored in a match and the goal difference, respectively, offering additional perspectives on match outcomes.
- **Result**: The outcome of the match (win, draw, loss) from the perspective of the home team, which could be useful for binary outcome analyses.

For the purpose of this study, we focused on matches from the Premier League era, starting from the 1992-1993 season, to ensure the relevance of our findings to contemporary football. The dataset was filtered to include only matches from the top tier of English football, aligning with Maher's original focus on first-division matches.

To calculate team-specific attacking and defensive strengths, a key component of Maher's model, we derived additional variables based on the average goals scored and conceded by each team within a given season. This approach allowed us to estimate each team's offensive and defensive capabilities, which are hypothesized to influence the number of goals scored in a match.

The data preparation process involved cleaning, filtering, and aggregating match data to construct a dataset suitable for Poisson regression analysis. This included handling missing values, ensuring consistency in team names across seasons, and calculating the aforementioned team strength metrics. The final dataset, comprising several thousand matches, provided a comprehensive basis for examining the dynamics of football scoring through statistical modeling, with an eye towards validating and extending Maher's pioneering work.

3 Methodology

To explore the statistical underpinnings of football scoring and the predictive power of Maher's model in contemporary football, we adopted a two-pronged approach. This section outlines the methodology used to calculate team strengths, model football scores, and assess the performance of different statistical models in predicting match outcomes.

3.0.0.1 Data Collection

Our analysis began with the collection of football match data, focusing on key variables such as match dates, teams, goals scored by home and away teams, and match outcomes. We utilized the `engsoccerdata` package in R (R Core Team 2022), which provides a comprehensive dataset of English football matches. This dataset served as the foundation for our analysis, offering a historical perspective on team performances across various seasons.

3.0.0.2 Calculation of Team Strengths

Recognizing the importance of team-specific attacking and defensive capabilities, we calculated the average goals scored and conceded by each team as proxies for their attacking and defensive strengths, respectively. These calculations were performed separately for home and away matches to account for the potential influence of home advantage. The following steps were taken:

1. **Attacking Strength**: For each team, the average number of goals scored in home and away matches was calculated for the current season, providing a measure of the team's offensive capability.
2. **Defensive Strength**: Similarly, for each team, the average number of goals conceded in home and away matches was computed, offering insight into the team's defensive performance.

These metrics were normalized to ensure comparability across different seasons and leagues, considering the varying levels of competition and scoring trends.

3.0.0.3 Statistical Modeling

With team strengths quantified, we proceeded to model football scores using the Poisson distribution, as suggested by Maher. Separate Poisson regression models were constructed for home and away goals, with the following general form:

- **Home Goals Model**: The number of goals scored by the home team was modeled as a function of the home team's attacking strength and the away team's defensive strength.
- **Away Goals Model**: The number of goals scored by the away team was modeled based on the away team's attacking strength and the home team's defensive strength.

These models allowed us to estimate the expected number of goals scored in a match, taking into account the relative strengths of the competing teams.

3.0.0.4 Model Evaluation and Comparison

To assess the adequacy of the Poisson model and explore potential improvements, we conducted goodness-of-fit tests and compared the Poisson model against alternative approaches, such as the Negative Binomial regression for overdispersed data and logistic regression for binary outcomes (win/loss). The evaluation criteria included:

- **Goodness-of-Fit**: Measured by comparing observed and expected frequencies of match outcomes.
- **Predictive Accuracy**: Assessed through cross-validation techniques, focusing on the model's ability to predict match scores and outcomes accurately.

- **Model Complexity**: Considered in the context of the trade-off between model simplicity and explanatory power.

Through this methodological framework, we aimed to provide a comprehensive analysis of football scoring dynamics, leveraging statistical models to uncover patterns and predict outcomes in the sport of football.

4 Results

4.0.0.1 Home Team Model Results

The model for predicting the number of goals scored by home teams highlighted the significant impact of both home attacking strength and away defensive strength (Table 1). Specifically, an increase in the home team's attacking strength was associated with a higher number of goals scored. Similarly, an increase in the away team's defensive strength paradoxically showed a positive correlation with home goals scored, a result that warrants further investigation to understand the underlying dynamics.

- **Home Attacking Strength**: For every unit increase in the home team's attacking strength, the expected number of goals scored by the home team increased by approximately 25%.
- **Away Defensive Strength**: Contrary to expectations, an increase in the away team's defensive strength was also associated with a slight increase in the expected number of goals scored by the home team.

4.0.0.2 Away Team Model Results

The away team model revealed a similar pattern, with both away attacking strength and home defensive strength playing significant roles in determining the number of goals scored by away teams (Table 2). The effect sizes were slightly smaller compared to the home team model, reflecting the general trend of lower scoring in away matches.

- **Away Attacking Strength**: An increase in the away team's attacking strength led to an increase in the expected number of goals scored by the away team, with a 21% increase for every unit increase in attacking strength.
- **Home Defensive Strength**: Interestingly, the model also indicated that stronger home defensive capabilities were associated with an increase in away goals scored, echoing the counterintuitive findings from the home team model.

4.0.0.3 Model Fit and Goodness-of-Fit Tests

The residual deviance for both models suggested a good fit to the data, although not perfect, indicating that while the Poisson regression model captures a significant portion of the variability in football scores, there are additional factors and complexities not accounted for by the model. The AIC (Akaike Information Criterion) values for both models provided a measure for comparing model fit, with lower values indicating a better fit relative to the complexity of the model.

5 Discussion

The results from our analysis suggest that team-specific attacking and defensive strengths, as derived from historical performance data, are significant predictors of the number of goals scored in Premier League matches. This finding aligns with Maher's original hypothesis regarding the importance of these factors in modeling football scores. However, the positive association between defensive strength and goals scored, for both home and away models, suggests that the relationship between team strengths and scoring may be more nuanced than initially anticipated.

These findings contribute to the ongoing discussion about the best approaches to modeling football scores and predicting match outcomes. While the Poisson regression model provides a valuable framework for understanding the dynamics of football scoring, the nuances and exceptions observed in the data highlight the potential need for more sophisticated models or additional explanatory variables to capture the full complexity of the sport.

```
Table: Table 1: Home Team Model Summary

|                        |   Estimate| Std. Error|   z value| Pr(>&#124;z&#124;)|
|:-----------------------|----------:|----------:|---------:|------------------:|
|(Intercept)             | -0.9224465|  0.0078968| -116.8123|                  0|
|home_attacking_strength |  0.2479412|  0.0022501|  110.1923|                  0|
|away_defensive_strength |  0.2528374|  0.0022330|  113.2281|                  0|
```

```
Table: Table 2: Away Team Model Summary

|                        |   Estimate| Std. Error|    z value| Pr(>&#124;z&#124;)|
|:-----------------------|----------:|----------:|----------:|------------------:|
|(Intercept)             | -1.1457977|  0.0099858| -114.74263|                  0|
|away_attacking_strength |  0.2127612|  0.0028559|   74.49990|                  0|
|home_defensive_strength |  0.2143541|  0.0028386|   75.51478|                  0|
```

6 Conclusion

Our exploration into the dynamics of football scoring, inspired by Maher's seminal 1982 study, has provided valuable insights into the applicability of statistical models for understanding and predicting outcomes in the Premier League. By employing a simplified version of Maher's Poisson regression model, we analyzed the relationship between team strengths—both attacking and defensive—and the number of goals scored in matches. Our findings underscore the significance of these factors in influencing match outcomes, consistent with Maher's original conclusions.

The analysis revealed that both home and away team strengths are crucial predictors of the number of goals scored, with attacking strength showing a positive correlation with goal-scoring, as expected. However, the observation that defensive strength also appeared to positively influence goal-scoring presents an intriguing anomaly, suggesting that the dynamics of football scoring may involve more complex interactions than those captured by the model.

Several key takeaways emerge from our study:

1. **Validation of Maher's Approach**: The continued relevance of Maher's model, as demonstrated by its applicability to modern Premier League data, highlights the enduring value of statistical methods in analyzing football. The Poisson distribution remains a robust framework for modeling football scores, reflecting the random nature of goal-scoring events.

2. **Complexity of Football Dynamics**: The counterintuitive findings regarding defensive strength indicate that football matches are influenced by a myriad of factors beyond the simplistic dichotomy of attacking and defensive capabilities. This complexity calls for further research, possibly incorporating more nuanced variables such as tactical formations, player-specific data, and situational factors like weather conditions or referee decisions.

3. **Future Directions for Research**: Our study opens the door for more sophisticated modeling approaches, including bivariate Poisson models to account for correlations between team scores, as suggested by Maher, and negative binomial regression to handle overdispersion in the data. Additionally, machine learning techniques could offer new avenues for capturing the complex interactions that influence football scoring.

4. **Practical Implications**: Beyond academic interest, understanding the statistical underpinnings of football scoring has practical implications for sports analysts, coaches, and betting markets. By refining predictive models, stakeholders can gain strategic insights, optimize team performance, and make informed decisions.

In conclusion, our journey through the statistics of football scoring, while inspired by Maher's pioneering work, has highlighted both the progress made in the field and the opportunities for further exploration. As football continues to evolve, so too will the methods we use to understand it, blending the passion for the sport with the precision of statistical analysis. The beautiful game, it seems, is as much a science as it is an art.

References

Curley, James. 2020. "Engsoccerdata: English and European Soccer Results 1871-2020."

Maher, Michael J. 1982. "Modelling Association Football Scores." *Statistica Neerlandica* 36 (3): 109–18.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.