Impact of Data Errors on Statisitcal Analyses*

Shivank Goel

March 19, 2024

This paper presents an analysis of the English football Championship matches, applying Poisson Regression models to understand the predictions and outcomes of the games. The research focuses on various influencing factors such as team performance, home-field advantage, and half-time scores. Data cleaning and selection processes have been applied to the dataset for effective modeling. By analyzing the number of goals scored the paper provides insights into match predictions and factors that significantly affect game results. This work contributes to sports analytics by enhancing strategies for teams and providing a predictive framework for analysts and coaches in the football industry.

1 Introduction

Football is not just a game rather it is a global phenomenon and to which millions are emotionally attached. At its core, it's a sport with a deep passion and the unpredictable nature of outcomes. In the 2024 English Championship, each match tells a story that is often reflected in the raw data of the game – goals scored, fouls committed, and shots taken. This study aims to study these narratives using Poisson Regression models, tools that are increasingly being used in sports analytics (Rob Mackenzie 2012)

Nowadays, knowing our stats can give teams and fans a big advantage. This paper discussed the exciting field of sports numbers by studying the data from football matches. It's not only interesting for people who like to predict match results, but it's also super useful for teams looking to get better and win more games.

This paper is structured as follows. In the Data Section, we denote how data was generated and processed. In the Model Section we dive into the findings we discovered following the model we built. In the Discussion Section, we address biases and weaknesses in the data that contribute to our findings, and how we approached.The last section is Conclusion and Acknowledgements.

*Code and data are available at: https://github.com/shivankgoel003/modelling_football_score

2 Data

The dataset for our analysis contains details from English football matches during the Championship 2024 season. It is taken from the football data UK portal (**uk?**), and offers a snapshot of the season, including match dates, participating teams, scores at full-time and half-time, and a variety of match statistics such as shots on target and fouls. To process and clean this data, we used the R programming language (R Core Team 2022). While cleaning the data, We identified duplicate entries for some matches. These were removed to prevent any skewing of results. Also missing data, particularly in player statistics and match outcomes, were filled using mean imputation where reasonable, ensuring a complete dataset for accurate modeling. A sample of cleaned data can be seen in Table 1.

Table 1: Sample of Cleaned Data

div	date	time	home_team	away_team	fthg	ftag	ftr	hthg	htag	htr	referee
E1	2023-08-04	20:00:00	Sheffield Weds	Southampton	1	2	A	0	1	A	R Madley
E1	2023-08-05	15:00:00	Blackburn	West Brom	2	1	H	2	0	H	D Whitestone
E1	2023-08-05	15:00:00	Bristol City	Preston	1	1	D	0	0	D	D Webb
E1	2023-08-05	15:00:00	Middlesbrough	Millwall	0	1	A	0	0	D	G Ward
E1	2023-08-05	15:00:00	Norwich	Hull	2	1	H	1	1	D	K Stroud
E1	2023-08-05	15:00:00	Plymouth	Huddersfield	3	1	H	1	1	D	M Donohue

3 Model

3.1 Model Set-up

In our study, we've implemented a thorough statistical analysis for football matches and predicted outcomes. We chose a Poisson Regression model as it can help to model count data, like goals scored in a match, considering the inherent discrete nature of such data. Poisson regression is used to analyze count data (e.g., the number of drinks per week; the number of arrests per year). Poisson regression is used to answer the questions such as what factors can predict the frequency of an event.(Wu and Little 2011)

The primary focus of our model was to estimate the effects of various factors on the number of goals scored by both home and away teams. Factors included team strength, historical performance, current season performance, home advantage, and match-specific variables like ball possession, number of shots on goal and player line-ups.

Our model can be described by the following equation:

$$\log(\lambda_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi}$$

Where: - λ_i represents the expected number of goals scored by team i. - $X_{1i}, X_{2i}, ..., X_{pi}$ are the explanatory variables including team strength, home advantage, etc. - $\beta_0, \beta_1, ..., \beta_p$ are the parameters of the model estimated from the data.

The model's validity was ensured through diagnostics and residual checks to confirm that the Poisson assumption held true. Overdispersion was checked to ensure the model's robustness, and alternative models, like Negative Binomial, were considered for comparative purposes.

3.2 Model Summary and Model Results

The results of the Poisson regression analysis are summarized in Table Table 2 below. Based on your model summary, the coefficients represent the log-relative change in the expected count of home goals (Full Time Home Goals, FTHG) for a one-unit change in the predictor, holding all other predictors constant. The standard errors of these coefficients are in parentheses.

The model included 171 observations from various matches and used these to understand how different factors affect the number of goals scored. The accuracy of the model, as shown by the Root Mean Square Error (RMSE), was 0.83. This number gives us a sense of how close our model's predictions are to the actual number of goals scored - the lower the RMSE, the more accurate the model is. In our case, the model has done a reasonably good job. Sachid Deshmukh (2019)

The AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) values provided us with a measure of the model's quality, considering both the number of predictors and the goodness of fit. While the values indicate that our model has room for improvement, it's a solid starting point for understanding match outcomes.

4 Conclusion and Results

From this analysis, we learn that while some factors have a clear influence on how many goals a team might score, the dynamics of football matches are complex. Some teams might perform differently at home, and in-match performance such as first-half goals and accurate shooting plays a critical role.

As with any model, ours isn't perfect. Yet, it's a valuable tool for teams and analysts looking to understand the game and plan their strategies accordingly.

References

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

Table 2: Poisson Regression Model Summary for Top English Football Teams

	Poisson Regression Model
(Intercept)	0.10
	(0.33)
HomeTeamBlackburn	−0.69
	(0.65)
HomeTeamBristol City	−0.72
	(0.66)
HomeTeamCardiff	−1.07
	(0.77)
HomeTeamCoventry	−0.28
	(0.59)
HomeTeamHuddersfield	−0.96
	(0.59)
HomeTeamHull	−0.61
	(0.58)
HomeTeamIpswich	−0.34
	(0.35)
HomeTeamLeeds	−0.42
	(0.35)
HomeTeamLeicester	−0.25
	(0.37)
HomeTeamMiddlesbrough	−0.70
	(0.59)
HomeTeamMillwall	−16.62
	(1212.04)
HomeTeamNorwich	−0.54
	(0.54)
HomeTeamPlymouth	−1.62
	(1.05)
HomeTeamPreston	−0.15
	(0.48)
HomeTeamQPR	−0.92
	(0.65)
HomeTeamRotherham	−0.64
	(0.59)
HomeTeamSheffield Weds	−1.10
	(0.77)
HomeTeamSouthampton	−0.27
	(0.36)
HomeTeamStoke	−1.85
	(1.05)
HomeTeamSunderland	−0.51
	(0.47)
HomeTeamSwansea	−0.82
	(0.59)
HomeTeamWatford	−0.44
	(0.54)
HomeTeamWest Brom	−0.30
	(0.36)
HTHG	0.27
	(0.07)

Rob Mackenzie, Christopher Cushion. 2012. "Performance Analysis in Football: A Critical Review and Implications for Future Research."https://www.researchgate.net/publication/233947861_Performa

Sachid Deshmukh, Vishal Arora, Michael Yampol. 2019. "Poisson Regression. Negative Binomial Regression. Multinomial Logistic Regression. Zero Inflated Poisson. Negative Binomial Regression." https://rpubs.com/myampol/Data621_Group4_HW5.

Wu, W., and T. D. Little. 2011. *Poisson Regression.* https://www.sciencedirect.com/topics/psychology/poisson-regression.