Poisson Regression Analysis: Exploring the Relationship Between Shots and Goals for Burnley as the Home Team*

Ping-Jen (Emily) Su

March 19, 2024

1 Introduction

The remainder of this paper is structured as follows. Section 2…. In this study, we explore goal scoring trends in football matches using a dataset obtained from https://www.football-data.co.uk/englandm.php. Inspired by Smith (2002) analysis of modelling association football scores but a simpler model. We aim to examine whether the number of goals scored increases as more shots are attempted during a match. This investigation is of interest as it can provide insights into team performance and strategies. We use R Core Team (2023) and Wickham et al. (2019) to help complete this paper.

2 Data

2.1 Dataset Overview

Given the topic we are interested, we can simulate the data shown in Figure 1

[*] Code and data are available at: https://github.com/emisu36/Premier-League-Analysis.

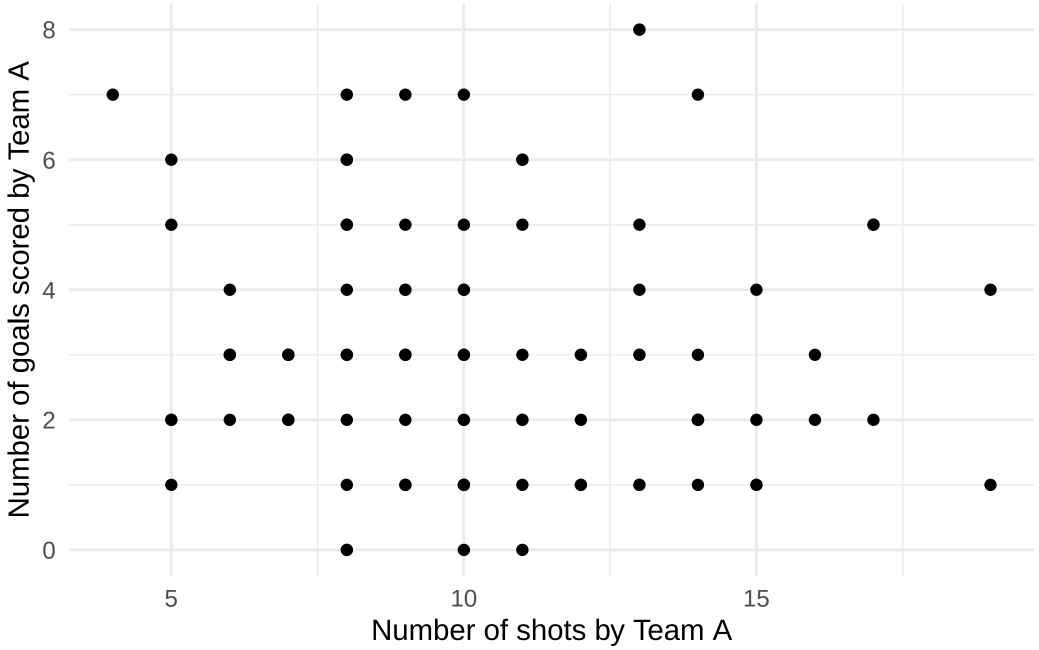

Figure 1: Simulation of the Data set

2.2 Dataset Preparation

We gather and prepare the data by downloading the match information from the provided source. The dataset is then cleaned and filtered to include relevant variables such as the number of shots and goals for each team in a match. In this time, we will be working to predict for when the home team is Burnley.

```
'data.frame':   15 obs. of  5 variables:
 $ Home     : chr  "Burnley" "Burnley" "Burnley" "Burnley" ...
 $ Away     : chr  "Man City" "Aston Villa" "Tottenham" "Man United" ...
 $ HomeGoals: int  0 1 2 0 1 0 1 5 0 0 ...
 $ AwayGoals: int  3 3 5 1 4 2 2 0 2 2 ...
 $ HS       : int  6 9 16 12 10 17 11 19 14 9 ...
```

2.3 Exploratory Data Analysis

We start by exploring the distribution of goals scored and shots attempted in the dataset. This includes calculating summary statistics and visualizing the relationship between these variables, as shown in Figure 2 .

```
`geom_smooth()` using formula = 'y ~ x'
```

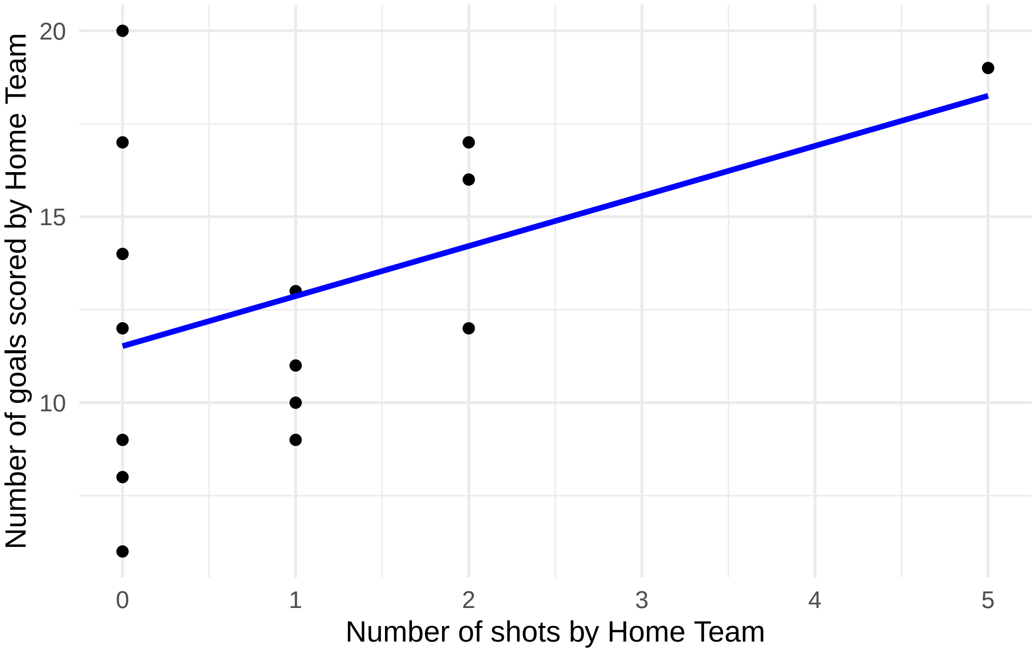

Figure 2: Simulation of the Data set

3 Model

3.1 Model set-up

Define y_i as the number of goals the team scored.

$$
\begin{aligned}
y_i|\lambda_i &\sim \text{Poisson}(\lambda_i) && (1)\\
\log(\lambda_i) &= \beta_0 + \beta_1 * \text{opponent}_i && (2)\\
\beta_0 &\sim \text{Normal}(0, 2.5) && (3)\\
\beta_1 &\sim \text{Normal}(0, 2.5) && (4)
\end{aligned}
$$

We run the model in R (R Core Team 2023) and used `glm` to form our model.

3.1.1 Model justification

Since poisson regression is commonly used to model count data when the outcome variable represents the number of times an event occurs within a fixed period of time or space. In football matches, goals scored by a team can be considered as count data, as they represent discrete events that happen over a fixed duration of time (the duration of the match). Additionally, the Poisson distribution is appropriate when the event rate is relatively low and events occur randomly and independently over time, which is a reasonable assumption for goals scored in football matches. Therefore, given that the outcome variable (goals scored by Burnley as the home team) meets the criteria for count data and the assumptions of the Poisson distribution hold reasonably well in this context, Poisson regression is a suitable choice for analyzing the relationship between the number of shots taken (predictor variable) and the number of goals scored (outcome variable) by Burnley as the home team.

4 Results

4.1 Goal Scoring Trends

To investigate the relationship between the number of shots attempted and goals scored, we fit a Poisson regression model. The model considers the number of shots as the explanatory variable and the number of goals as the response variable.

```
Call:
glm(formula = HomeGoals ~ HS, family = poisson, data = cleaned_data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.00620    1.04053  -1.928   0.0538 .
HS           0.14279    0.06631   2.153   0.0313 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 24.412  on 14  degrees of freedom
Residual deviance: 19.485  on 13  degrees of freedom
AIC: 42.807

Number of Fisher Scoring iterations: 5
```

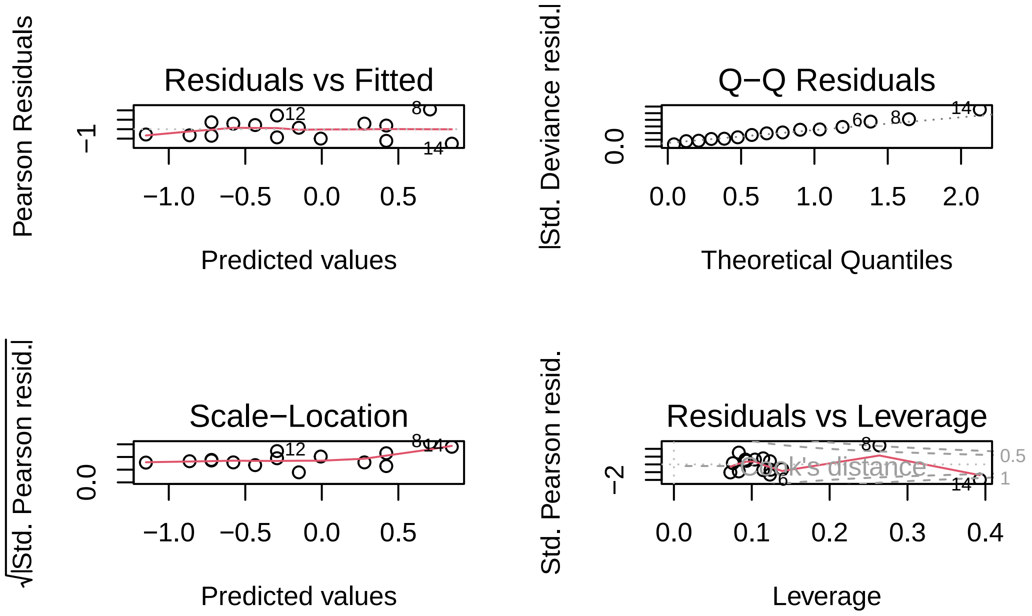

Figure 3: Diagnostic Plots

The Poisson regression model was employed to analyze the relationship between the number of shots (HS) and the number of goals scored by the home team (HomeGoals) in football matches. The model yielded the following results:

Coefficients:

Intercept: The estimated intercept (β_0) is -2.00620 with a standard error of 1.04053. The z-value associated with the intercept is -1.928, yielding a p-value of 0.0538, suggesting marginal significance at the 0.05 level. HS: The estimated coefficient ((β_1)) for the number of shots (HS) is 0.14279, with a standard error of 0.06631. The z-value for this coefficient is 2.153, corresponding to a p-value of 0.0313, indicating significance at the 0.05 level.

Model Fit:

Null Deviance: The null deviance, representing the difference between the null model (with no predictors) and the observed data, is 24.412 on 14 degrees of freedom. Residual Deviance: The residual deviance, measuring the difference between the fitted model and the observed data, is 19.485 on 13 degrees of freedom. AIC (Akaike Information Criterion): The AIC value for the model is 42.807.

5 Discussion

5.1 Interpretation

With the results in Figure 3, the intercept term (-2.00620) represents the expected number of goals scored by the home team when the number of shots (HS) is zero. However, since zero shots are not practically feasible in a football match, the interpretation of the intercept in this context may not be meaningful.

The coefficient for the number of shots (HS) is 0.14279, indicating that for each additional shot taken by the home team, the expected number of goals scored increases by approximately 0.143, holding all other variables constant. This coefficient is statistically significant (p = 0.0313), suggesting that there is evidence to reject the null hypothesis of no relationship between the number of shots and the number of goals scored by the home team.

5.2 Model Evaluation:

The model's AIC value of 42.807 suggests that, among competing models, the current Poisson regression model provides a relatively good balance between model fit and complexity.

5.3 Discussion and Conclusion:

The results indicate that there is a positive association between the number of shots taken by the home team and the number of goals scored, supporting the intuitive notion that a higher volume of shots increases the likelihood of scoring goals. However, it's essential to consider other factors, such as the quality of shots, defensive strategies of the opposing team, and various situational factors, which may also influence goal-scoring outcomes in football matches.

Overall, the findings contribute to understanding the factors influencing goal scoring in football matches and provide insights for coaches, analysts, and decision-makers to optimize team strategies and performance. Further research could explore additional variables and more complex models to enhance the predictive accuracy of goal-scoring trends in football.

References

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Smith, Richard L. 2002. "A statistical assessment of Buchanan's vote in Palm Beach County." *Statistical Science* 17 (4): 441–57. https://doi.org/10.1214/ss/1049993203.

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.