Goal Patterns in European Football: A Statistical Analysis of Match Outcomes*

Chenyiteng Han

March 19, 2024

This study delves into the statistical patterns of goal scoring in European football matches, drawing on an extensive dataset from the European Soccer Database available on Kaggle. By employing a range of R packages designed for data manipulation, cleaning, and statistical modeling, this research examines the relationship between team identifiers and match outcomes. Initial Poisson regression models revealed significant overdispersion, prompting a transition to Negative Binomial regression to achieve a more accurate portrayal of the count data. This paper elucidates the methodological choices between logistic, Poisson, and Negative Binomial regressions, ultimately justifying the use of the latter due to its capability to handle overdispersed data. The results offer new insights into home and away goal scoring trends, contributing to the broader understanding of team performance in European leagues.

1 Introduction

The realm of competitive football is rich with data, each match being a confluence of strategy, skill, and chance that leaves behind a trail of statistics ripe for analysis. At the heart of this investigation is the quest to uncover the patterns of goal scoring and understand the nuances that distinguish one match from another. This paper presents a comprehensive analysis of European football matches, employing advanced statistical methods to dissect the intricacies of the game's most definitive outcome: the goal.

Leveraging the data science capabilities of R, we accessed and manipulated a detailed dataset from the publicly available European Soccer Database on Kaggle. The dplyr package, part of the powerful tidyverse suite, afforded us an efficient and intuitive means of cleaning and preparing our data for analysis. This process involved selecting key variables that shed light

*Code and data from this analysis are available at: https://github.com/Hhnxxxxxx/European-Football.git

on the number of goals scored by both home and away teams, allowing us to probe into the potential home-field advantages and team performance dynamics.

Our statistical journey began with Poisson regression models, which are well-suited for modeling count data such as goals. However, our rigorous diagnostic tests indicated the presence of overdispersion, leading us to pivot to Negative Binomial regression models. These models, equipped to handle the extra variance observed in our data, painted a more accurate and nuanced picture of goal-scoring trends.

As we navigate through the intricate statistical landscape, this paper discusses the rationale behind our choice of modeling techniques. We articulate why logistic regression, typically reserved for binary outcomes, was not a suitable contender for our count data analysis. In contrast, the Negative Binomial regression emerged as the superior choice, adeptly accommodating the overdispersion that the Poisson regression models could not.

With the adoption of the Negative Binomial regression models, we enhance the robustness of our findings, ensuring that our inferences are firmly rooted in the realities of the data. This study not only contributes valuable insights to the domain of sports analytics but also serves as a testament to the power of statistical modeling in extracting meaningful stories from raw numbers.

2 Data

The analysis of match outcomes within the comprehensive European Soccer Database, openly available on Kaggle, was undergirded by the R software environment and a suite of its packages, offering an extensive toolkit for data science (R Core Team (2022)). Direct interactions with the SQLite database were facilitated through the RSQLite package (Wickham and Müller (2022)), providing a seamless database management experience in R. Data manipulation and cleaning were adeptly handled using the tidyverse package, a collection of R packages designed for data science that simplifies many common data handling tasks (Wickham et al. (2019)). The dplyr package, an integral part of the tidyverse (Wickham et al. (2019)), delivered a powerful and user-friendly syntax for data manipulation, while broom (Robinson and Hayes (2020)) elegantly converted statistical analysis outputs into tidy data frames, making them amenable to further analysis and interpretation. For more advanced statistical modeling, especially for count data, the MASS package (Venables and Ripley (2002)) supplied the necessary functions to accurately model overdispersion through negative binomial regression. The documentation and reporting process was augmented by knitr (Xie (2014)), integrating R code with prose to produce dynamic reports. Finally, the kableExtra package (Zhu (2022)) was utilized to enhance the knitr package's kable function, enabling the creation of sophisticated and aesthetically pleasing tables that effectively communicate the results of the statistical models.

The dataset under review has been sourced from the extensive European Soccer Database, which is publicly available on Kaggle. This particular subset of data has been subjected

to a cleansing process to refine the contents for analytical purposes. The original, expansive database includes a variety of tables encompassing detailed information about football matches across European leagues.

2.1 Data Cleaning Process

The cleaning process involved the selection of relevant columns that capture the essential details of the matches, leading to a focused and streamlined dataset for subsequent analysis. Specifically, the columns retained through the process are as follows:

- `id`: A unique identifier assigned to each match, facilitating easy referencing and data integrity.
- `season`: The campaign during which the match was played, spanning from the 2008/2009 season to the 2015/2016 season, providing a temporal dimension to the data.
- `home_team_api_id`: The unique API identifier for the home team, delineating the team playing on their home turf.
- `away_team_api_id`: The unique API identifier for the away team, distinguishing the team playing outside their home venue.
- `home_team_goal`: The total number of goals netted by the home team during the match, offering insights into the offensive strength and home advantage.
- `away_team_goal`: The total number of goals netted by the away team, presenting a measure of the team's performance in an away setting.

2.2 Purpose and Significance of Selected Variables

The rationale behind selecting these specific columns was to distill the dataset to capture the match's key aspects that are most relevant for outcome analysis. The focus on the number of goals scored by home and away teams enables investigations into phenomena such as home advantage and comparative team performance.

The variable `season` adds temporal context, allowing for the analysis across different seasons and observation of evolutionary trends in the data. The identifiers `home_team_api_id` and `away_team_api_id` are crucial for distinguishing between teams, linking match results to specific teams for comprehensive analyses when cross-referenced with other datasets. The quantitative measures `home_team_goal` and `away_team_goal` are fundamental for modeling match scores, analyzing the effectiveness of team strategies, and predicting outcomes in future matches.

The sample table of the cleaned dataset is shown in Table **??**.

Table 1: Sample of the Cleaned Dataset

ID	Season	Home Team API ID	Away Team API ID	Home Team Goal	Away Team Goal
1	2008/2009	9987	9993	1	1
2	2008/2009	10000	9994	0	0
3	2008/2009	9984	8635	0	3
4	2008/2009	9991	9998	5	0
5	2008/2009	7947	9985	1	3
6	2008/2009	8203	8342	1	1

3 Model

In this analysis, we explore the relationship between football match outcomes—specifically, the number of goals scored by home and away teams—and the teams' unique identifiers.

3.1 Poisson Regression Models

We constructed two Poisson regression models—one for the home team goals and another for the away team goals. Poisson regression is suitable for count data, such as goals scored, which follows a Poisson distribution. The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time since the last event.

3.1.1 Home Team Goals Model

The model for home team goals is formalized as follows:

$$\log(\mu_{\text{home}}) = \beta_0 + \beta_1 \times \text{HomeTeamAPIID} + \beta_2 \times \text{AwayTeamAPIID}$$

Where: - μ_{home} represents the expected count of goals scored by the home team. - β_0 is the intercept of the model. - β_1 and β_2 are the coefficients for the home and away teams' API IDs, respectively.

3.1.2 Away Team Goals Model

Similarly, the model for away team goals is given by:

$$\log(\mu_{\text{away}}) = \alpha_0 + \alpha_1 \times \text{HomeTeamAPIID} + \alpha_2 \times \text{AwayTeamAPIID}$$

Where: - μ_{away} is the expected count of goals scored by the away team. - α_0 is the intercept. - α_1 and α_2 are the coefficients corresponding to the home and away teams' API IDs.

In both models, the `home_team_api_id` and `away_team_api_id` serve as predictors, and their coefficients measure the association of each team's identity with the number of goals scored. The `log` link function relates the linear predictors to the expected log count of goals.

These models provide insights into the scoring patterns associated with different teams and may reveal the presence of any home-field advantage or away-field disadvantage.

The summary of both models is shown in Table **??** and Table **??**.

Table 2: Summary of Poisson Regression Model for Home Teams

term	estimate	std.error	statistic	p.value
(Intercept)	0.4411598	7.17e-03	61.528855	0.0000000
home_team_api_id	-0.0000011	4.00e-07	-2.669400	0.0075987
away_team_api_id	0.0000004	3.00e-07	1.172712	0.2409112

Table 3: Summary of Poisson Regression Model for Away Teams

term	estimate	std.error	statistic	p.value
(Intercept)	0.1501320	0.0081957	18.318368	0.0000000
home_team_api_id	0.0000010	0.0000004	2.708559	0.0067576
away_team_api_id	-0.0000011	0.0000005	-2.432691	0.0149871

3.2 Overdispersion in Poisson Regression Models

In the application of Poisson regression models, a fundamental assumption is that the mean and variance of the count data are equal. This characteristic is intrinsic to the Poisson distribution, which is defined by a single parameter governing both its mean and variance. Nonetheless, empirical data frequently violate this assumption, exhibiting what is known as overdispersion, where the observed variance surpasses the expected mean.

3.2.1 Diagnostic Tests for Overdispersion

To diagnose overdispersion within our Poisson regression models for home and away team goals, we conducted a series of tests:

- We calculated the Pearson residuals for each model, which measure the difference between the observed and predicted counts, normalized by the predicted standard deviation.
- We determined the variance of these Pearson residuals. For a model fitting the data well, we would anticipate this variance to approximate 1. A variance significantly greater than 1 signals overdispersion.
- We computed a Chi-squared goodness-of-fit test for each model. Under the correct model, the sum of squared Pearson residuals should follow a Chi-squared distribution. A large Chi-squared statistic indicates a poor fit, suggesting that the model's assumptions may not hold for the data.

3.2.2 Test Results and Implications

The test results shown in Table **??** revealed a residual variance exceeding 1 for both the home and away goal models, coupled with exceedingly low p-values from the Chi-squared tests. This indicates a substantial deviation from the Poisson model's assumptions, signifying the presence of overdispersion.

3.2.3 Transitioning to a Negative Binomial Regression Model

Given the evidence of overdispersion, it is statistically prudent to transition to a Negative Binomial regression model. This model extends the Poisson regression by introducing an additional parameter to explicitly model overdispersion, permitting the variance to outstrip the mean. Adopting the Negative Binomial regression framework enhances the model's flexibility, offering a more accurate fit for count data characterized by overdispersion and leading to more reliable inferential statistics.

Table 4: Summary of Poisson Regression Model for Away Teams

	Residuals		Chi-squared Test		
Model	Residuals_Var	Squared_Residuals_Mean	Chi_squared	P_value	Overdispersion
Home	1.089075	1.089033	28291.99	0	TRUE
Away	1.123344	1.123301	29182.23	0	TRUE

3.3 Negative Binomial Regression Models

With the detection of overdispersion in the Poisson regression models, we have proceeded to adopt Negative Binomial regression models for both home and away team goals. This approach is more suitable for data where the variance exceeds the mean, which is often the case in count data such as goals scored in football matches.

Negative Binomial regression is a type of generalized linear model that generalizes Poisson regression by introducing an extra parameter to account for the overdispersion. The probability distribution of the Negative Binomial model allows for the variance to be greater than the mean, which provides a better fit for overdispersed data.

3.3.1 Home Team Goals Model

The modified model for home team goals accounts for the overdispersion and is defined by the following relationship:

$$\log(\mu_{\text{home}}) = \beta_0 + \beta_1 \times \text{HomeTeamAPIID} + \beta_2 \times \text{AwayTeamAPIID} + \text{Overdispersion}$$

Here, μ_{home} is the expected count of goals scored by the home team, adjusted for overdispersion. β_0 is the model's intercept, while β_1 and β_2 are the coefficients for the home and away team API IDs. The term "Overdispersion" is a placeholder for the Negative Binomial model's dispersion parameter.

3.3.2 Away Team Goals Model

For the away team goals, the Negative Binomial model is similarly adjusted:

$$\log(\mu_{\text{away}}) = \alpha_0 + \alpha_1 \times \text{HomeTeamAPIID} + \alpha_2 \times \text{AwayTeamAPIID} + \text{Overdispersion}$$

In this equation, μ_{away} represents the adjusted expected count of goals for the away team. α_0 is the intercept, and α_1 and α_2 correspond to the home and away team API IDs. The dispersion parameter is included to account for the extra variance observed in the data.

The API IDs of the teams serve as predictors in these models, with their coefficients providing a measure of the teams' scoring propensity, all the while adjusting for the variability beyond what is captured by the Poisson assumption.

By fitting these Negative Binomial models, we aim to obtain more reliable estimates that accommodate the observed overdispersion, thereby enhancing the robustness of our inferences regarding team performances.

The summary of both models is shown in Table **??** and Table **??**.

Table 5: Summary of Negative Binomial Regression Model for Home Teams

term	estimate	std.error	statistic	p.value
(Intercept)	0.4411820	0.0074704	59.057429	0.0000000
home_team_api_id	-0.0000011	0.0000004	-2.581898	0.0098259
away_team_api_id	0.0000004	0.0000004	1.117974	0.2635781

Table 6: Summary of Negative Binomial Regression Model for Away Teams

term	estimate	std.error	statistic	p.value
(Intercept)	0.1501501	0.0086780	17.302289	0.0000000
home_team_api_id	0.0000010	0.0000004	2.523693	0.0116129
away_team_api_id	-0.0000011	0.0000005	-2.323770	0.0201378

4 Discussion

In our investigation of football match outcomes, particularly the count of goals scored, logistic regression was considered but ultimately not selected as the primary analytical approach. Logistic regression is typically utilized for binary or categorical outcome variables, such as win/loss or goal/no goal scenarios. However, our research focus required modeling the actual number of goals, which is a count variable. While logistic regression could potentially be used to model the probability of scoring any number of goals, it does not naturally account for the distribution of counts, nor does it easily extend to incorporate the variability seen in the number of goals scored across matches. This limitation becomes particularly pronounced when addressing the issue of overdispersion—a phenomenon not adequately managed by logistic regression. Poisson and negative binomial regression models are specifically designed for count data and include the flexibility to handle overdispersion, thus providing a more suitable and precise modeling framework for our data structure and research objectives.

Initially, we implemented Poisson regression models due to their appropriateness for count data, which assumes the mean and variance of the distribution to be equal — an assumption intrinsic to the Poisson distribution. However, our diagnostic tests suggested the presence of overdispersion in the data, as evidenced by the Pearson residuals' variance being significantly greater than 1 and the resulting large chi-squared values.

The overdispersion indicated that the variability in our data was too great to be adequately modeled by the Poisson distribution. As a consequence, the standard errors estimated by the Poisson models would be underestimated, potentially leading to incorrect inferences. To address this, we turned to the Negative Binomial regression model, which introduces an additional parameter to model the overdispersion explicitly. This model allows the variance to exceed the mean, providing a more flexible fit for our overdispersed count data.

Furthermore, while logistic regression is suitable for binary outcomes, our response variables — the number of goals scored by home and away teams — are counts, making logistic regression less appropriate in this context. Negative Binomial regression is more apt for our analysis since it can handle the count nature of the response variable along with the overdispersion.

In conclusion, the choice to employ Negative Binomial regression was driven by the need to account for the extra-Poissonian variation observed in our data. By fitting these models, we achieved a more reliable and nuanced understanding of the factors influencing goal-scoring in European football matches, ensuring our statistical inferences are robust and reflective of the underlying data patterns.

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/.

Robinson, David, and Alex Hayes. 2020. "Broom: An r Package for Converting Statistical Analysis Objects into Tidy Data Frames." *The R Journal* 12 (1): 57–66. https://doi.org/10.32614/RJ-2020-048.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with s*. New York: Springer.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, and et al. 2019. *Tidyverse: Easily Install and Load the Tidyverse*. https://CRAN.R-project.org/package=tidyverse.

Wickham, Hadley, and Kirill Müller. 2022. *RSQLite: 'SQLite' Interface for r*. https://CRAN.R-project.org/package=RSQLite.

Xie, Yihui. 2014. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Zhu, Hao. 2022. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.