From Touches to Goals: Modeling the Efficacy of Possession in Football*

Evaluating Poisson vs. Negative Binomial Models in Predicting Home Team Success

Siqi Fei

March 18, 2024

This study delves into the intricate relationship between a football team's ball possession, as indicated by the number of times players touch the ball, and the team's scoring success. Through rigorous statistical analysis, we compared Poisson and negative binomial regression models to assess which better predicts the final score based on possession data. Our findings revealed a nuanced connection where more ball possession does not always equate to higher scores, challenging common football assumptions. The implications of this research extend beyond sports analytics, providing insights into the complex dynamics of performance metrics and their interpretation, thereby enriching the strategic understanding of team sports.

Table of contents

*Code and data are available at: https://github.com/FXXFERMI/Modelling_association_football_scores.git.

1 Introduction

Football is the world's most popular sport, with a vast audience and professionals across the globe. Germany, with its reputation for football excellence, hosts the Bundesliga every year. Bundesliga is a league that draws global attention due to its high level of competition and skill. The data from this league is highly regarded for the precision and dependability, making it ideal for in-depth analysis. In this study, we examine data from the Bundesliga season 2023-2024 to better understand the outcomes of matches, particularly how ball possession relates to scoring goals.

This study uses data from Kaggle for the year 2024 (Orcun 2024), applies Poisson and negative binomial regression models to closely examine the link between how often home team players have the ball and the number of goals they score. The aim is to see how keeping the ball influences the team's ability to score.

Echoing Maher's article (Maher 1982), which instead of rejecting the Poisson model for football scores, further explored it by considering how well teams attack and defend, our study revisits these models with newest Bundesliga data. We're comparing the Poisson and the Negative Binomial models to figure out which one better reflects the number of goals in current football games. Through our analysis with the latest data, we aim to uncover which model truly captures the scoring trends in football today

Our paper outlines the importance of analyzing match data, how we did our study, what we found, and why it's important. We show how our findings help us better understand football and its scores. Our research offers new insights for coaches, players, and anyone interested in the science behind the sport.

2 Data

For this paper, we used data from the Kaggle 2024 Bundesliga Complete Season Analysis (Orcun 2024), which we worked with in R (R Core Team 2023), a language for statistical computing. The `tidyverse` suite (Wickham et al. 2019), with its various packages like `ggplot2` (Wickham 2016), `dplyr` (Wickham et al. 2023), `readr` (Wickham, Hester, and Bryan 2024), and `tibble` (Müller and Wickham 2023), made handling the data easier and more precise. We summarized our data results using `knitr` package (Xie 2023) and model results using the `modelsummary` package (Arel-Bundock 2022). The `here` package (Müller 2020) helped keep our files organized and our analysis reproducible.

2.1 Data Measurement

In our study, the measurement of data plays a pivotal role in understanding the dynamics of football matches within the Bundesliga. We focus on three main data points: Home Team, Final Score of the Home Team and Number of Times the Home Team Players Touched the Ball. There are 63 observations in the raw dataset, before analyzing, we thoroughly clean our data by removing non-essential information. 59 observations remains in our study.

The Figure 1 presents a statistical overview of the average number of times each Bundesliga team's players touched the ball at home games (denoted as "avg_touch") along with the variance of these touches (denoted as "var_touch"). These two factors provide insights into the teams' playing style and control of the game when they are on familiar ground. A higher average suggests a team that often maintains possession, potentially indicating a dominant play style. Conversely, high variance can imply inconsistency in possession, suggesting that the team's control over the ball fluctuates significantly from game to game.

Home	avg_touch	var_touch
Augsburg	528.7500	5398.2500
Bayern Munich	747.6667	294.3333
Bochum	531.0000	12168.0000
Darmstadt 98	589.6667	29912.3333
Dortmund	772.5000	5168.3333
Eint Frankfurt	752.0000	8172.0000
Freiburg	593.0000	1971.0000
Heidenheim	505.6667	3397.3333
Hoffenheim	560.3333	6666.3333
Köln	590.0000	2887.0000
Leverkusen	834.2500	31148.2500
M'Gladbach	622.0000	11306.0000
Mainz 05	540.3333	4358.3333
RB Leipzig	664.7500	16950.9167
Stuttgart	717.5000	4349.6667
Union Berlin	598.0000	15138.0000
Werder Bremen	529.6667	17814.3333
Wolfsburg	625.0000	1521.0000

Figure 1: The Mean and Variance of Number of Times the Home Team Players Touched the Ball

The Figure 2 shifts focus to the outcomes of the home team's efforts, showcasing the mean and variance of the final scores across the same span of seasons. The mean of the final scores (denoted as "avg_score") offers a direct indicator of a team's offensive effectiveness, while

the variance(denoted as "var_score") gives us an understanding of their scoring consistency. Teams with a high average score are generally seen as strong offensive teams, and those with low variance in scores are typically more predictable in their scoring performance. Together, these statistics paint a picture of the attacking prowess and reliability of each team in the Bundesliga during home matches.

Home	avg_score	var_score
Augsburg	2.2500000	1.5833333
Bayern Munich	4.0000000	7.0000000
Bochum	1.0000000	0.0000000
Darmstadt 98	2.6666667	2.3333333
Dortmund	2.0000000	2.0000000
Eint Frankfurt	1.0000000	0.6666667
Freiburg	1.6666667	0.3333333
Heidenheim	2.3333333	2.3333333
Hoffenheim	1.6666667	1.3333333
Köln	0.6666667	0.3333333
Leverkusen	3.7500000	0.9166667
M'Gladbach	0.7500000	0.9166667
Mainz 05	0.6666667	0.3333333
RB Leipzig	2.5000000	4.3333333
Stuttgart	4.0000000	1.3333333
Union Berlin	2.0000000	8.0000000
Werder Bremen	2.0000000	4.0000000
Wolfsburg	2.0000000	0.0000000

Figure 2: The Mean and Variance of Final Score of the Home Team

2.2 Data Distribution

Figure 3 visualizes the relationship between the number of times the home team players touched the ball and the final score achieved by the home team in a selection of football matches. From the distribution of data points, it appears that there isn't a straightforward linear correlation between these two factors. While we might expect more ball touches to correlate with more goals, the data does not show a consistent trend supporting this.

The points are relatively scattered, suggesting that while ball possession (as indicated by touches) is an important aspect of the game, it does not directly guarantee more goals. This could indicate that other factors, such as the quality of touches, team strategy, defensive strength of the opposition, and effectiveness in the final third, also play critical roles in converting possession into goals.

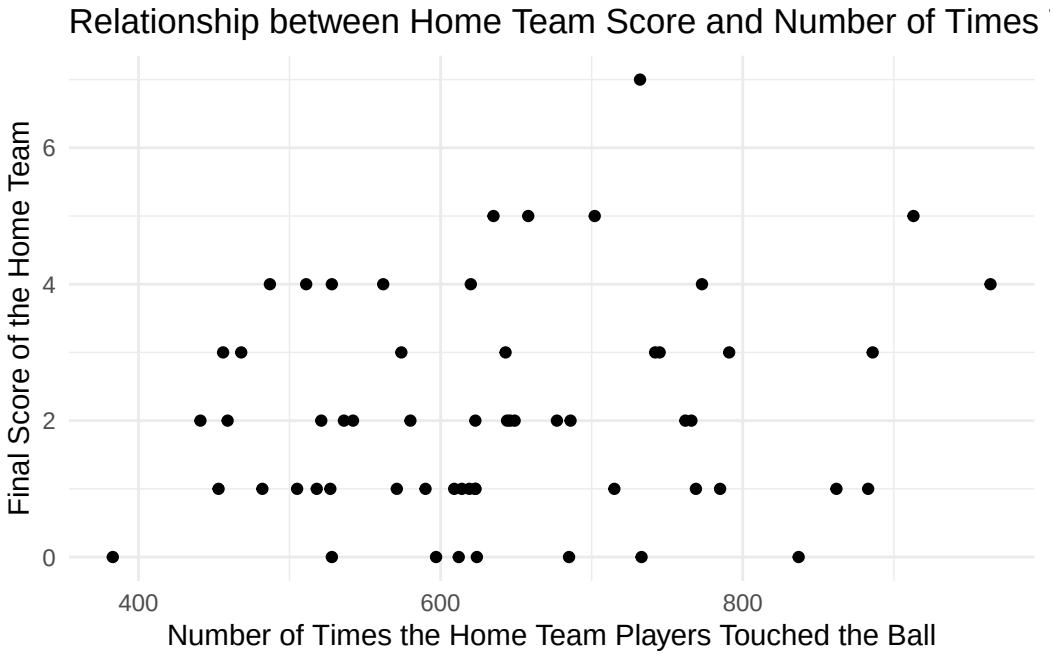


Figure 3: The Relationship between Home Team Score and Number of Times Touched the Ball

3 Model

3.1 Model set-up

We run the model in R (R Core Team 2023) using the `rstanarm` package of (Goodrich et al. 2022). We use the default priors from `rstanarm`.

3.2 Model Application and Findings

In our analysis, we applied both Poisson and negative binomial regression models to explore the relationship between the number of times the home team players touched the ball and the final score of the home team. Table 3 presents the estimated coefficients from both models, which are remarkably similar, indicating that each additional touch of the ball is associated with a very slight increase in the expected number of goals scored.

Table 3: Modeling the relationship between Number of Times the Home Team Players Touched the Ball and Final Score of the Home Team

	Poisson	Negative Binomial
(Intercept)	0.147	0.147
Touches_x	0.001	0.001
Num.Obs.	59	59
Log.Lik.	−104.952	−104.952
ELPD	−107.1	−107.1
ELPD s.e.	5.5	5.5
LOOIC	214.2	214.2
LOOIC s.e.	11.0	11.0
WAIC	214.1	214.1
RMSE	1.55	1.55

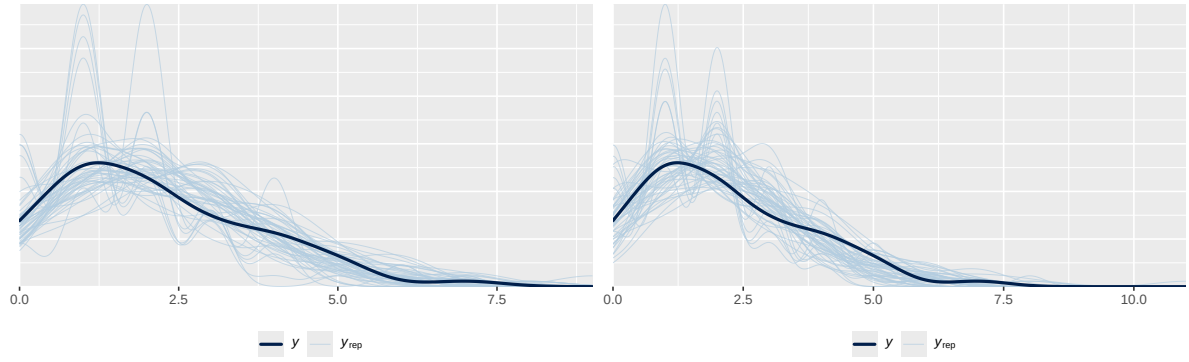
3.3 Comparative Model Performance

The similarities between the Poisson and negative binomial models, as shown in Table 3, also extend to the metrics used to evaluate model fit and predictive accuracy—namely, the Log Likelihood, ELPD (Expected Log Pointwise Predictive Density), LOOIC (Leave-One-Out Cross-Validation Information Criterion), WAIC (Watanabe-Akaike Information Criterion), and RMSE (Root Mean Square Error). All these metrics suggest nearly identical performance of the two models on our dataset, with no significant differences.

3.4 Posterior Predictive Checks

Figure 4 compares the posterior predictive checks for both models. These checks allow us to visually assess how well our models' predictions match the observed data. The overlap of the predicted values (in blue) around the observed values (in black) appears similar across both models, with neither showing a distinct advantage in capturing the distribution of the observed scores.

The choice between using a Poisson or a negative binomial regression often comes down to the data's dispersion. The Poisson regression assumes the mean and variance of the count data are equal (equidispersion), while the negative binomial model allows for variance to exceed the mean (overdispersion), which is often more realistic in real-world scenarios such as football scoring. Despite this theoretical difference, our analysis indicates that both models performed similarly on our specific dataset, perhaps due to a lack of overdispersion in the data or other dataset-specific characteristics.



(a) Poisson model

(b) Negative binomial model

Figure 4: Comparing posterior prediction checks for Poisson and negative binomial models

4 Result

When choosing the best model to analyze football match data, like the number of times players touch the ball and how many goals they score, we have to look closely at how spread out the scores usually are. The Poisson model is the simpler one and is based on the idea that the average number of goals scored is around the same as how much the scores vary from game to game. However, real match scores can often vary a lot more than that, which is why the negative binomial model is usually thought to be more flexible—it can handle data where the scores vary widely.

Our study put these two models to the test using Bundesliga match data to see if more ball touches by the home team really lead to more goals. What we discovered was that touching the ball more often didn't really increase the team's chances of scoring by much, no matter which model we used. What was unexpected was that the negative binomial model, which many analysts and researchers prefer for football data, didn't give us better results than the Poisson model. This was surprising because it didn't seem to matter that the scores varied—both models ended up being equally useful.

Our research gives us new insights into predicting football scores and suggests that the difference between the Poisson and negative binomial models might not be as clear-cut as we thought. This finding challenges some usual beliefs in sports analytics and shows how important it is to choose the right model based on the actual data we have.

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

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

Maher, M. J. 1982. "Modelling Association Football Scores." *Statistica Neerlandica* 36 (3): 109–18. https://doi.org/https://doi.org/10.1111/j.1467-9574.1982.tb00782.x.

Müller, Kirill. 2020. *here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Müller, Kirill, and Hadley Wickham. 2023. *tibble: Simple Data Frames*. https://CRAN.R-project.org/package=tibble.

Orcun, Memocan. 2024. "Bundesliga 2021-2024 Complete Season Analysis." https://www.kaggle.com/datasets/memocan/bundesliga-2021-2024-complete-season-analysis.

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

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

——— et al. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.