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

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

Siqi Fei

April 2, 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.

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 (**bundesliga_dataset_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.

*Code and data are available at: LINK.

Echoing Maher's article (**Maher1982?**), 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

Some of our data is of penguins (Figure 1), from Horst, Hill, and Gorman (2020).

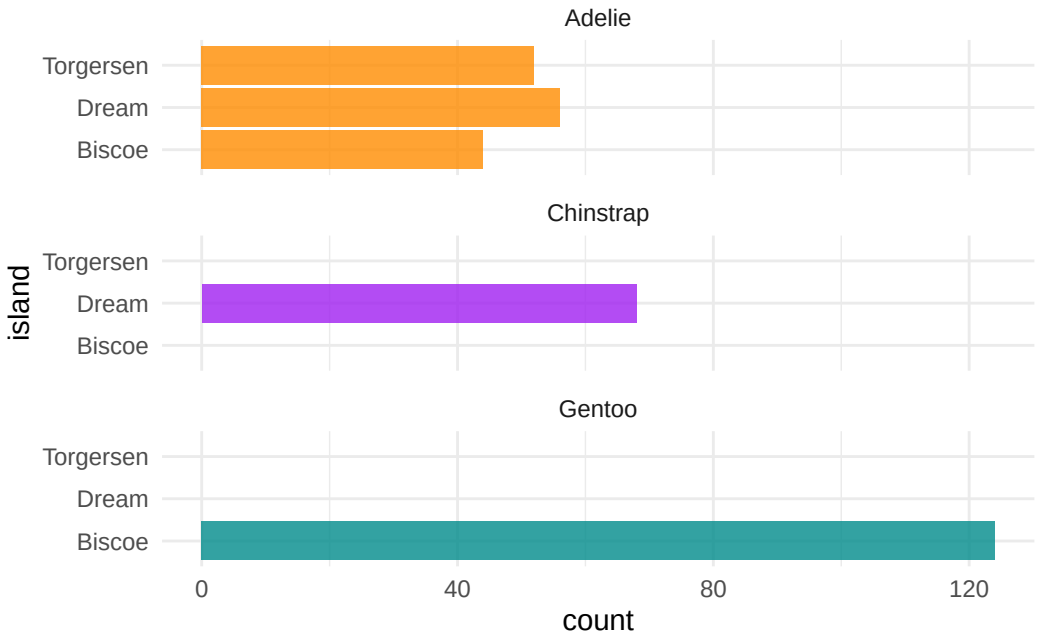


Figure 1: Bills of penguins

Talk more about it.

And also planes (Figure 2). (You can change the height and width, but don't worry about doing that until you have finished every other aspect of the paper - Quarto will try to make it look nice and the defaults usually work well once you have enough text.)

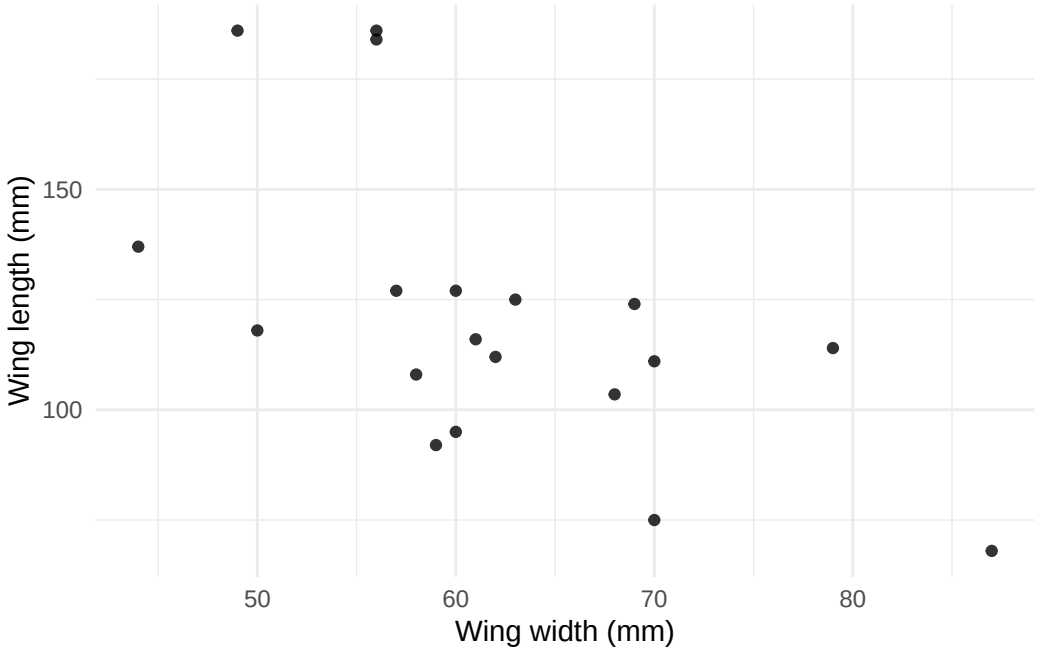


Figure 2: Relationship between wing length and width

Talk way more about it.

3 Model

The goal of our modelling strategy is twofold. Firstly,...

Here we briefly describe the Bayesian analysis model used to investigate... Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \alpha + \beta_i + \gamma_i \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\gamma \sim \text{Normal}(0, 2.5) \tag{5}$$
$$\sigma \sim \text{Exponential}(1) \tag{6}$$

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`.

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular...

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in Table 1.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Table 1: Explanatory models of flight time based on wing width and wing length

	First model
(Intercept)	1.12
	(1.70)
length	0.01
	(0.01)
width	−0.01
	(0.02)
Num.Obs.	19
R2	0.320
R2 Adj.	0.019
Log.Lik.	−18.128
ELPD	−21.6
ELPD s.e.	2.1
LOOIC	43.2
LOOIC s.e.	4.3
WAIC	42.7
RMSE	0.60

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

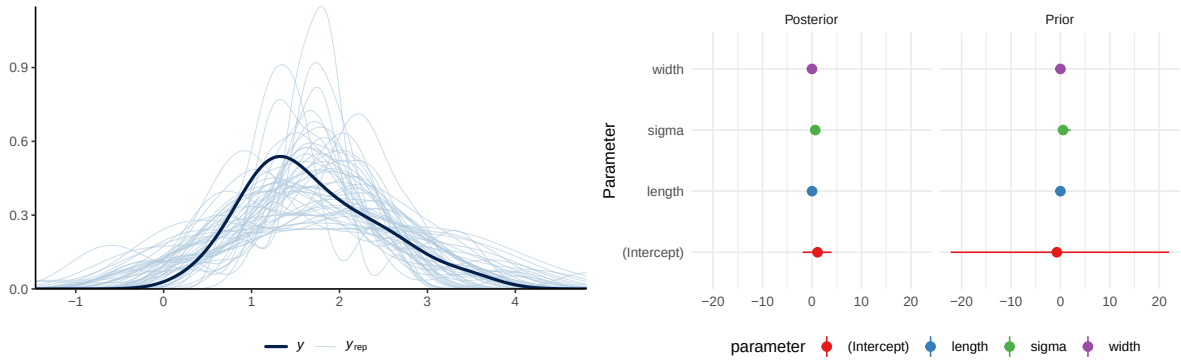
In Figure 3a we implement a posterior predictive check. This shows...

In Figure 3b we compare the posterior with the prior. This shows...

B.2 Diagnostics

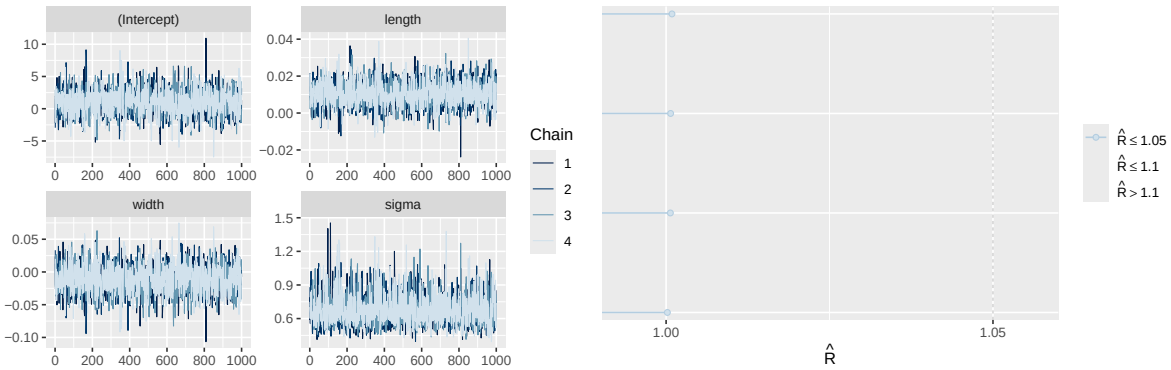
Figure 4a is a trace plot. It shows... This suggests...

Figure 4b is a Rhat plot. It shows... This suggests...



(a) Posterior prediction check (b) Comparing the posterior with the prior

Figure 3: Examining how the model fits, and is affected by, the data



(a) Trace plot (b) Rhat plot

Figure 4: Checking the convergence of the MCMC algorithm

References

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

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. *Palmerpenguins: Palmer Archipelago (Antarctica) Penguin Data*. https://doi.org/10.5281/zenodo.3960218.

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