A Quasi-Poisson Approach Goals in Football, What contributes to them?

Nayani Jani
Christopher Odoom
Denis Folitse
Animesh Sengupta

University of Massachusetts, Amherst

November 27, 2022

Background of study

Football has become a part of the world and is one of the biggest source of entertainment in the world. It has created job and opportunities to people all over the world. It's important in this modern time cannot be overlooked.There are five major football leagues in the world; English premier league, German BundesLiga, Spanish Laliga, Italian Seria A and France League one. Football supporters of English primier league all over the world always claim that, their league is the difficult one among all the five major leagues, is this claim true?. Football fans all over the world have insufficient answer to this question. This study tries to provide answer to this question.

Problem statement

Argument: If a league has a lot of goals scored in a season it is deemed not competitive because of the quality of opponents (vice versa as well). Goal: Does the factor league determine the amount of goals combined with other variables? If not, what predictors are associated with the response Goals?

Objectives

This study is directed by the specific objectives stated below;

1. Determine if there is a significant difference in goals across the 5 major leagues.
2. Determine if there are other predictors that are associated with the goals other than the different leagues.

Methodology

This study used a data which was collected from `https://fbref.com`, We used the 2021-2022 season statistics for EPL, La Liga, Bundesliga, Ligue 1 and Serie A in the analysis. The data comprise of 98 observations and 18 variables. Models we considered includes Poisson regression, Quasi - Poisson Regression. Diagnostic Tests: Test for Overdispersion, Residual Vs. Fitted Plot, Normal QQ-Plot

Poisson Regression

This is type a regression used when the response variable is a COUNT variable. It models the logarithm of the expected value of occurrence of an event as a linear function of some predictor variables.
The mathematical equation for the model is given below:

$$log(E(no.Goals|X)) = \beta_0 + \beta^{'} X$$

Alternatively, we can write it in an exponential form as:

$$E(No.Goals|X) = e^{(\beta_0 + \beta^{'} X)}$$

The method used here to estimate the paramters is the Maximum Liklihood estimator.

Assumptions of poisson Regression

- The number of goals must be scored in a fixed time.
- $$E(no.Gls|X) = Var(no.Gls|X)$$

Estimation of parameter

The estimation method used here is the maximum likelihood estimation method. The procedure is briefly described below: Given the model:

$$\lambda(x) = E(Y|X) = e^{\beta' X}$$

Here Y, the response follow a Poisson distribution, therefore we have

$$P(y|x, \beta') = \frac{\lambda(x)^y e^{-\lambda(x)}}{y!}$$

$$= \frac{e^{y\beta' X} e^{-e^{\beta' X}}}{y!}$$

Now suppose we have data set (X, y) with n rows, then given an set of parameters β the likelihood is given by

$$L(\beta|X, Y) = \prod_{i=1}^{n} \frac{e^{y_i \beta' X_i} e^{-e^{\beta' X_i}}}{y_i!}$$

The goal is to find β which maximize this likelihood.

Quasi-Poisson

This study used this variant of Poisson regression because, the data is under-dispersion(the conditonal variance is smaller than expected). This type of model add a parameter called the dispersion which help solve the problem of under-dispersion. This reason for this model is to produce credible inference. The dispersion parameter is given by

$$\hat{\phi} = \frac{1}{n-K} \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i}$$

where K is the number of parameters.
such that

$$Var(Y|X) = \hat{\phi} E(Y|X)$$

if $\hat{\phi} = 1$ we have Possion regression
if $\hat{\phi} > 1$ we have over-dispersion
if $\hat{\phi} < 1$ we have under-dispersion

Pearson Residuals

This is a useful measure for diagnosing the appropriateness of a Poisson regression. This type of residual is a weighted residual and when ploted with the fitted points tells us if the Poisson regression is ok and there is no issue with dispersion. This measure is computed with the formula below.

$$PearsonResidual = \frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i}}$$

Results

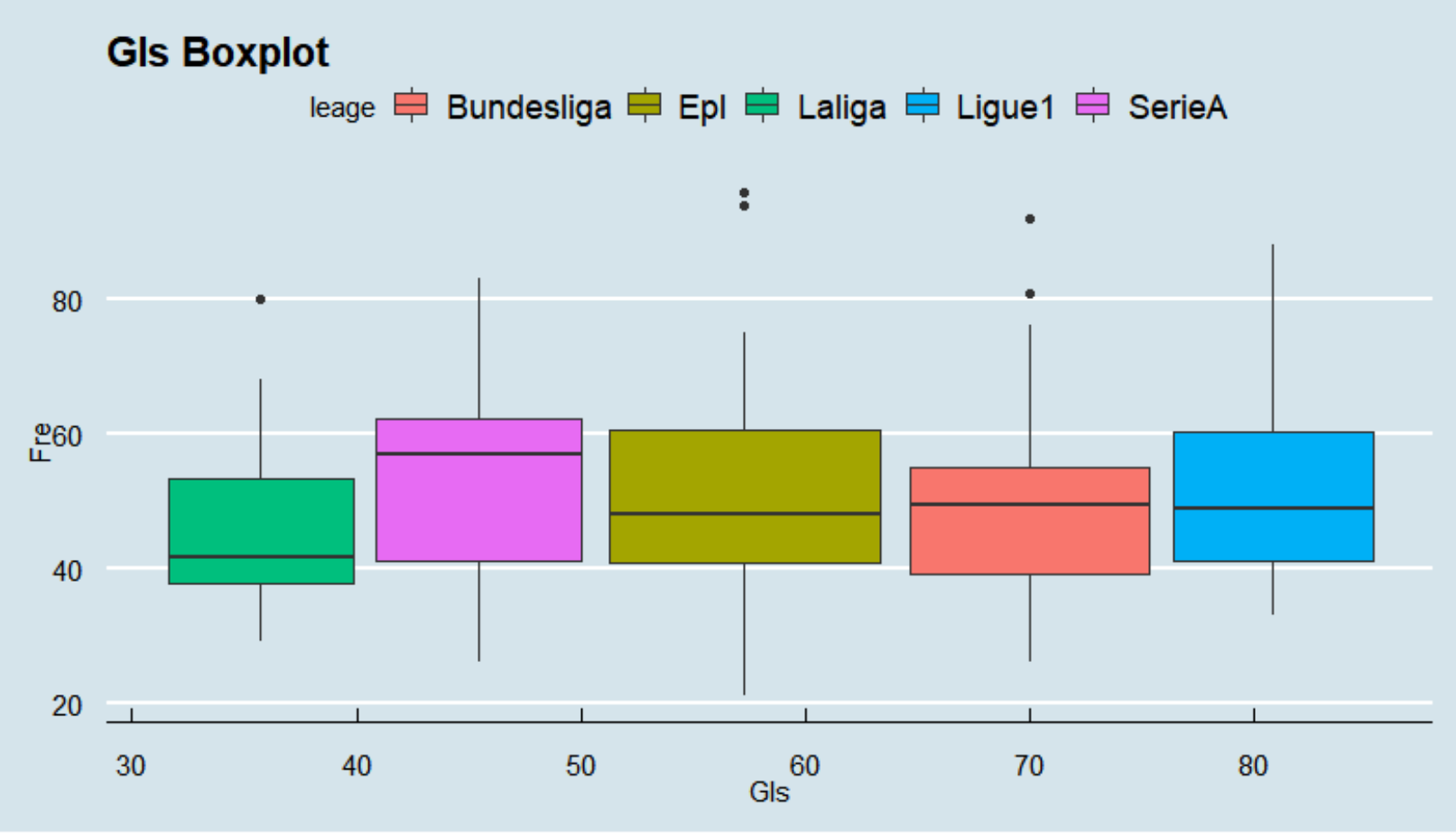

Figure1: The plot show the goals distribution across the 5 Leagues

Results