A Quasi-Poisson Regression Analysis:

Goals in Football, What contributes to them?

Nayan Jani
Christopher Odoom
Denis Folitse
Animesh Sengupta

University of Massachusetts, Amherst

December 2, 2022

Outline

Background of study

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..
- There are five major football leagues in the world; English premier league, German BundesLiga, Spanish Laliga, Italian Serie A and France League one.
- Football supporters in the various leagues all over the world always claim that, their league is the difficult one among all the five major leagues, is this claim true?.

Motivation

Motivation

Arguments:

- If teams score lot of goals week in week out in a league, that leagues quality is brought to question. Goes as far as labeling the league a "Farmers League"
- Goal scoring stats has been used a lot of late to determine how good an attacking player performs.
- Unfortunately, Due to this labeling of some leagues as having a low quality, Some players are still undermined no matter the number of goals they bang week in week out.
- Other times, defenders are blamed to be too poor.

Objectives

Objectives

Main Objective:

- Does the type of league determine the number of goals scored per season combined with other variables?

Specific Objectives:

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

Data

- We used data from the 2021-2022 season for five major leagues: EPL, La Liga, Bundesliga, Ligue 1 and Serie A in the analysis.
- The data comprise of 98 observations and 17 variables.
- The data was pulled from `https://fbref.com`

Variable Description

Data

Type	Description
Gls	Number of Goals
Team	Name of Team
NumAtt	Number of Attacking Players (G+A >9)
League	Name of League
no.sh	Number of Shots
SoT	Shots on Target
PK	Number of Penalty Kicks
FK	Number of Free Kicks

Att. 3rd.T	Attacking 3rd Touches
Succ.Drib	Successful Dribbles
Att.Drib	Dribbles Attempted
Touches	Number of Touches
Prog.T	Progressive Passes Received
GCA.drib	Successful Dribbles that lead to a Goal
ShortAT_Pass	Short Attacking Pass
MediumAT_Pass	Medium Attacking Pass
LongAT_Pass	Long Attacking Pass

Data collected from https://fbref.com/, 2021-2022 season for EPL, La Liga, Bundesliga, Ligue 1 and Serie A

Poisson Regression

In Deciding for the model;

- **Random Component:** The random component identifies the response variable and selects a probability for it. For our objectives, the response variable, No. of Goals is a COUNT variable that follows a Poisson distribution.

- **Systematic component:** In here, this component help specifies the explanatory variables as a linear predictor. For each explanatory predictor, i, the linear predictors are given as;

$$\beta_0 + \beta_i' X_i$$

- **Link Function:** We specify the function, $g(.)$ that connects the random and systematic component. i.e relating our expected response variable to the linear predictors. For our response: Since the expected No. of goals cannot be negative; we use the log link.

This give us the mathematical equation below:

$$\log(E(no.Goals|X)) = \beta_0 + \beta_i' X_i$$

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

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

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

Assumptions of Poisson Regression:

- The response, no. of goals must be a count data.
- 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.

Model Diagnostic

To make sure the assumption

$$E(no.Gls|X) = Var(no.Gls|X)$$

is met, we used the dispersion parameter. This 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

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

This implies that if:

- $\hat{\phi} = 1$ then the assumption is met and then we have Possion regression
- $\hat{\phi} > 1$ we have over-dispersion
- $\hat{\phi} < 1$ we have under-dispersion

Pearson Residuals

- We also did a Pearson residual to diagnose/verify if the poission regression is appropriate.

- This weights the residuals and when plotted with the fitted points helps us determine if there is an issue with dispersion or not. This measure is computed with the formula below.

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

Quasi-Poisson

As the assumption is violated, what next?

- In this study we used a variant of Poisson regression called Quasi-Poisson Regression for our final model as a result of the violation of the assumption:

$$E(no.Gls|X) = Var(no.Gls|X)$$

- The data is under-dispersed(the conditional variance is smaller than expected).
- This model factors in the dispersion parameter which help solve the problem of under-dispersion. **The reason for this model is to produce credible inference.**

Pearson Residuals

- The Pearson residual corrects for the unequal variance in the raw residuals by dividing by the standard deviation. The formula for the Pearson residuals is

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

Results

Histogram

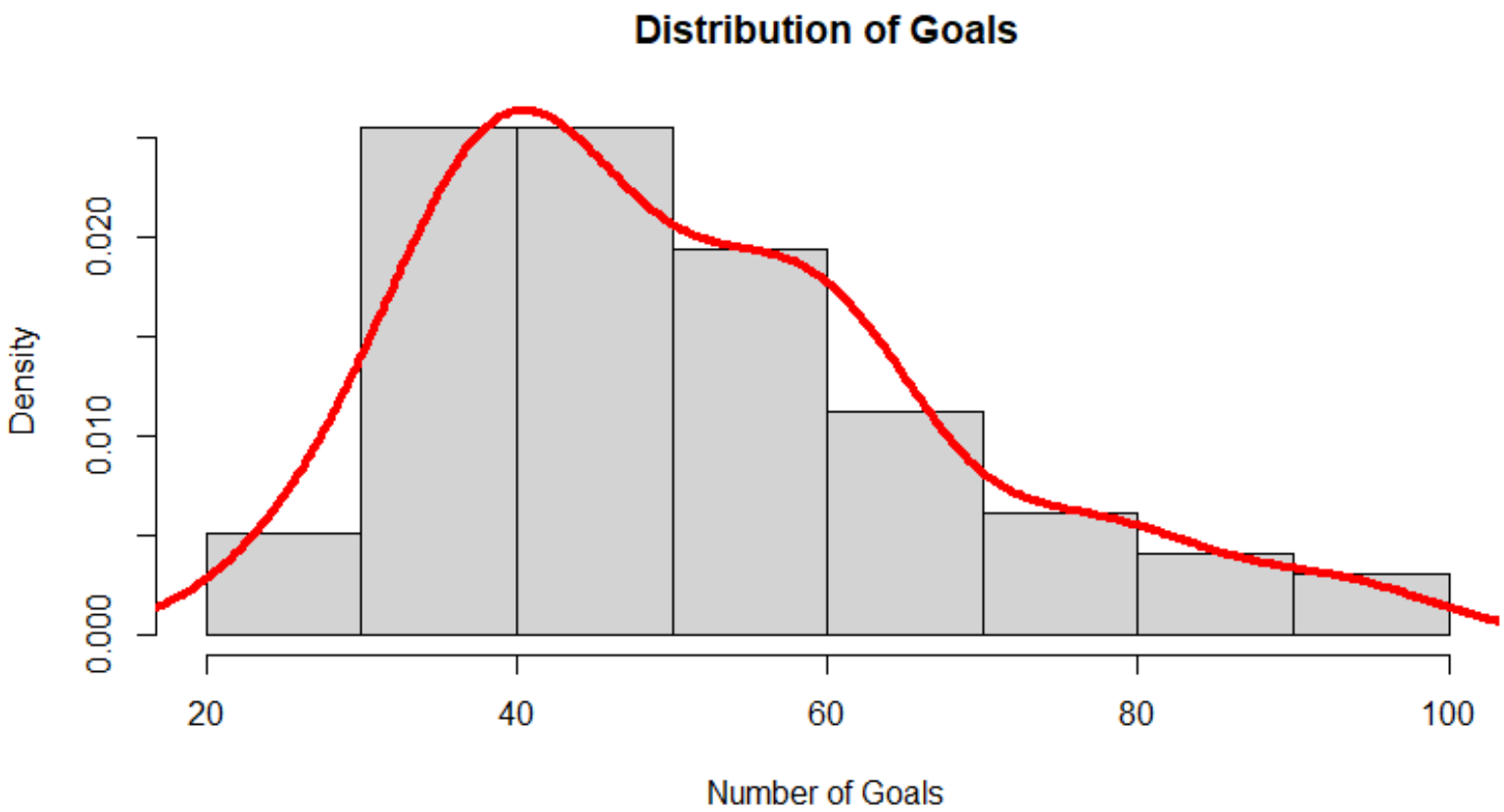

Figure1: The plot show the goals distribution across the 5 Leagues

Boxplot

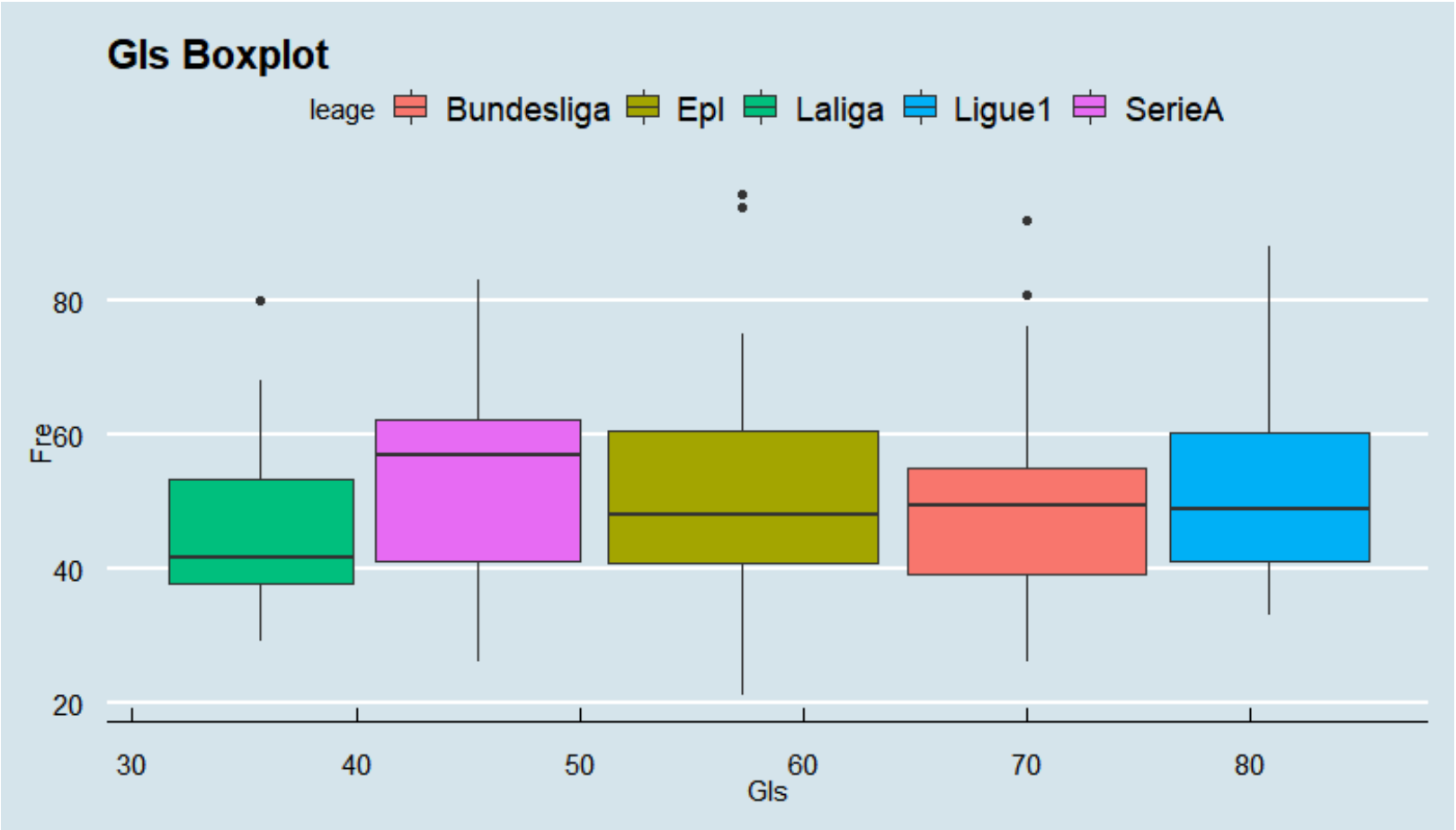

Figure1: The plot show the goals distribution across the 5 Leagues

Linear Regression

Table: Naive OLS Model (Multiple $R^2 = 0.9045$)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-14.943	6.097	-2.451	0.016
leageBundesliga	5.375	2.059	2.610	0.011
leageLaliga	-0.487	1.872	-0.260	0.795
leageLigue1	1.169	1.895	0.617	0.539
leageSerieA	-0.614	1.995	-0.308	0.759
No.PI	0.133	0.166	0.802	0.425
SoT	0.296	0.034	8.672	0
PK	0.755	0.273	2.771	0.007
FK	-0.147	0.100	-1.471	0.145
Att.3rd.T	0.001	0.001	0.710	0.479
Succ.Drib	-0.022	0.012	-1.785	0.078
ShortAT_Pass	0.001	0.001	2.102	0.039
NumAtt	3.271	0.626	5.229	0.00000

Results

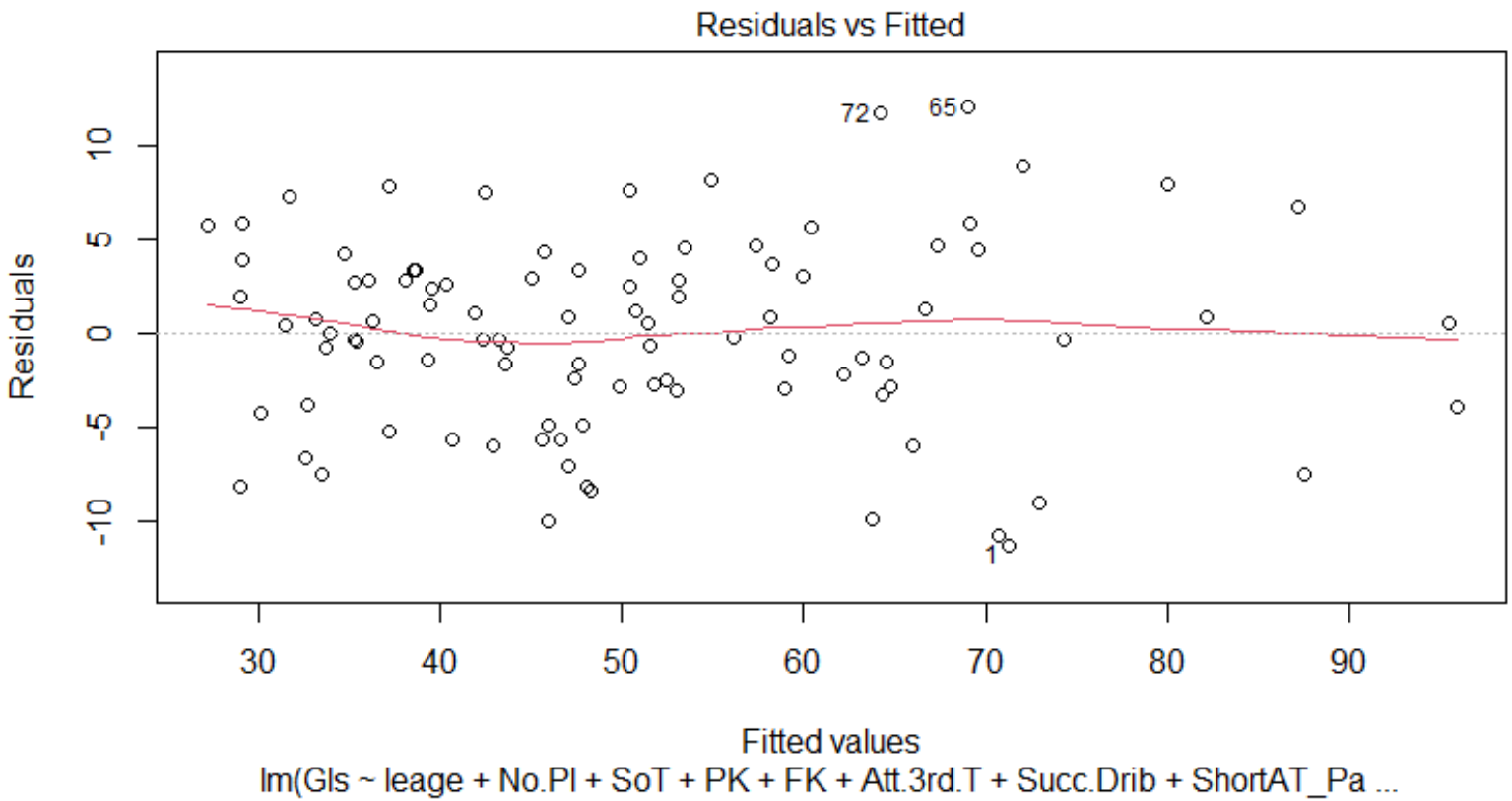

Figure3: The residual plots for OLS

Results

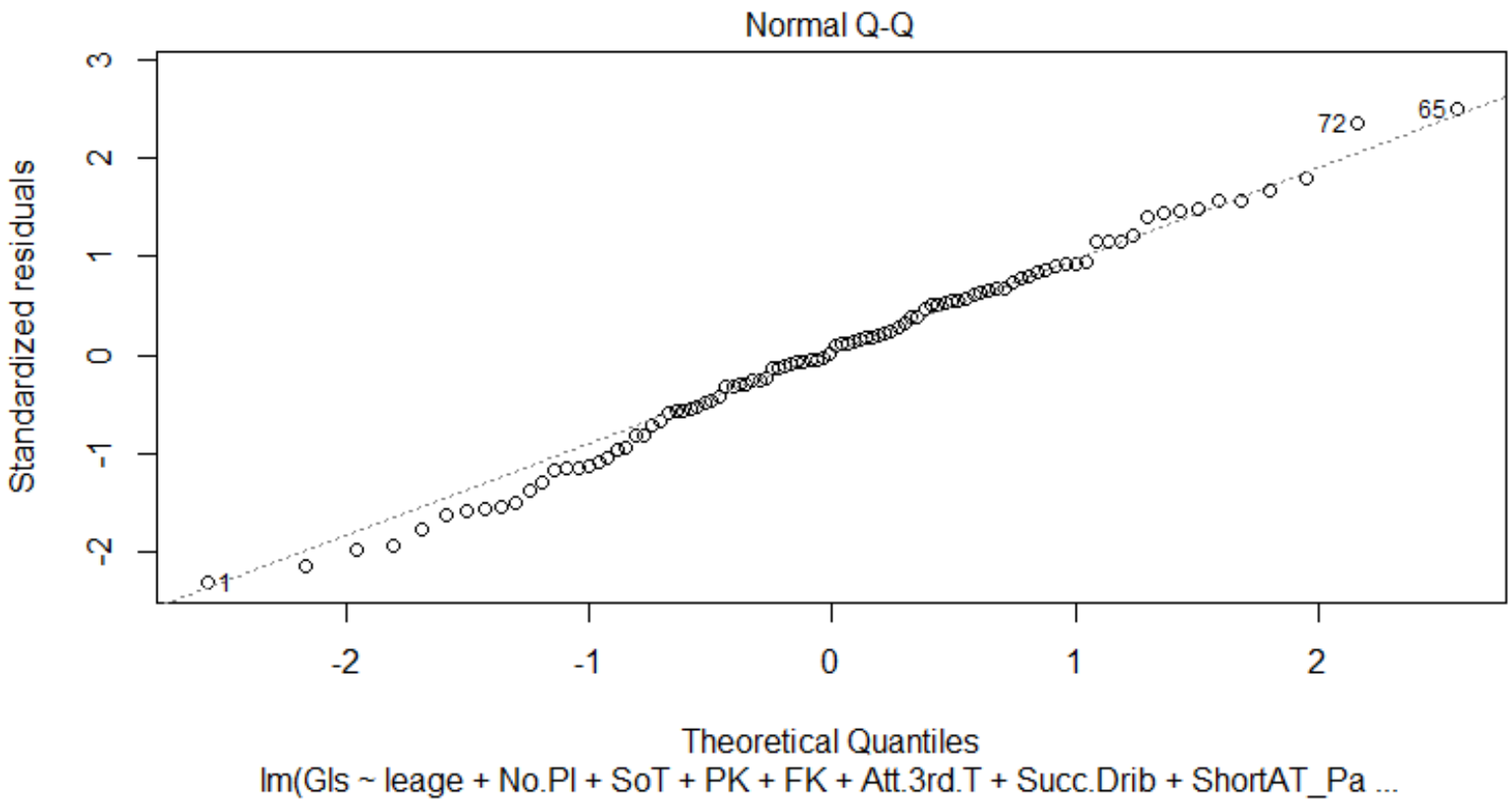

Figure4: The QQ plot for OLS

Poisson Regression Output

Table: Poisson Regression Output

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.853	0.103	27.592	0
leageBundesliga	0.093	0.053	1.770	0.077
leageLaliga	-0.007	0.048	-0.138	0.890
leageLigue1	0.036	0.048	0.742	0.458
leageSerieA	0.004	0.048	0.089	0.929
SoT	0.005	0.001	6.503	0
PK	0.016	0.007	2.244	0.025
FK	-0.003	0.003	-0.993	0.321
Att.3rd.T	-0.00001	0.00003	-0.283	0.777
Succ.Drib	-0.0005	0.0003	-1.655	0.098
ShortAT_Pass	0.00003	0.00002	1.885	0.059
NumAtt	0.073	0.015	4.719	0.00000

Dispersion parameter = 1 Residual Deviance = *ResidualDeviance*= 58.267

Residual Plot

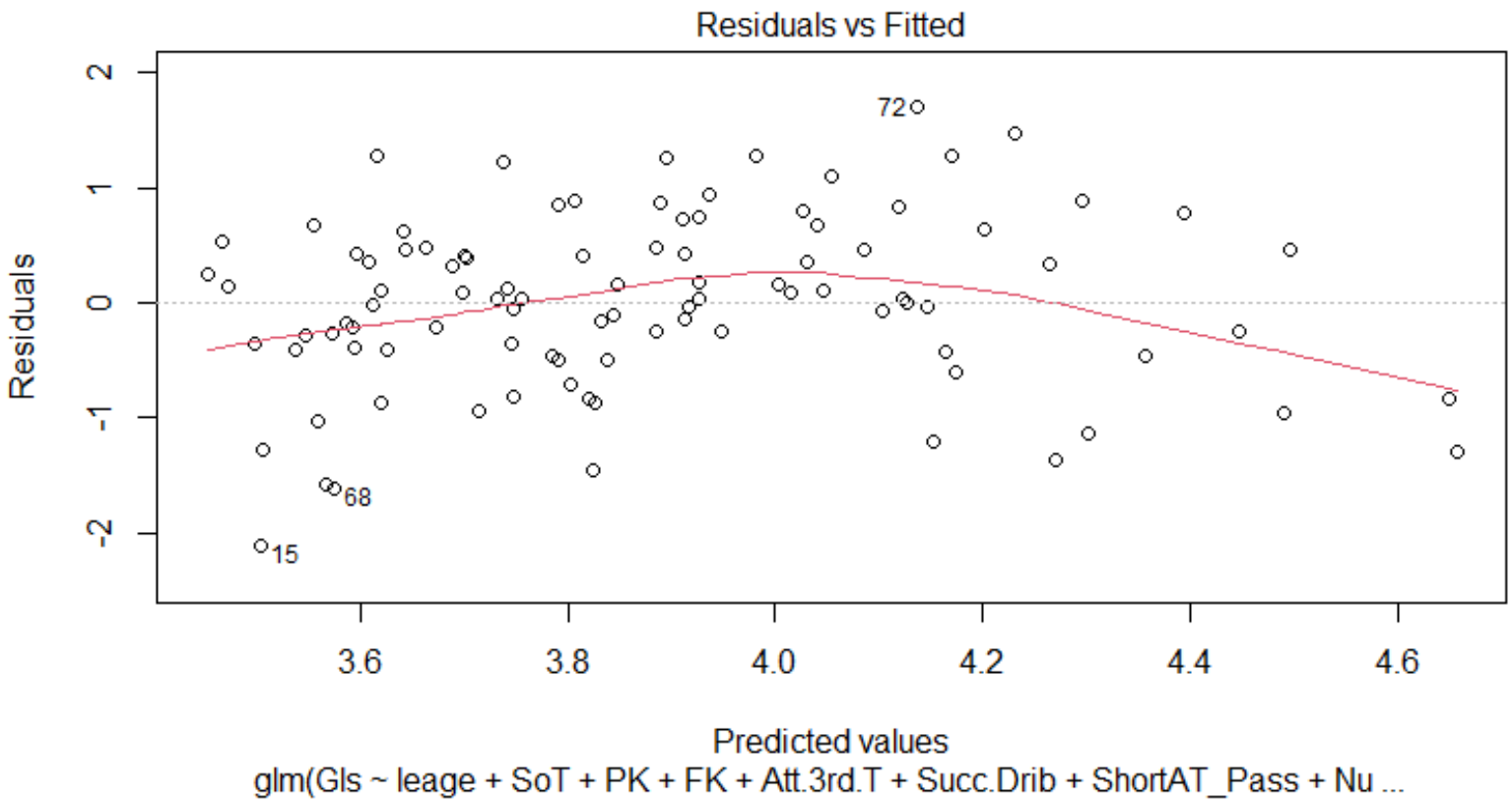

Figure5: Residual Plot for Poisson Regression

Pearson Residual Plot

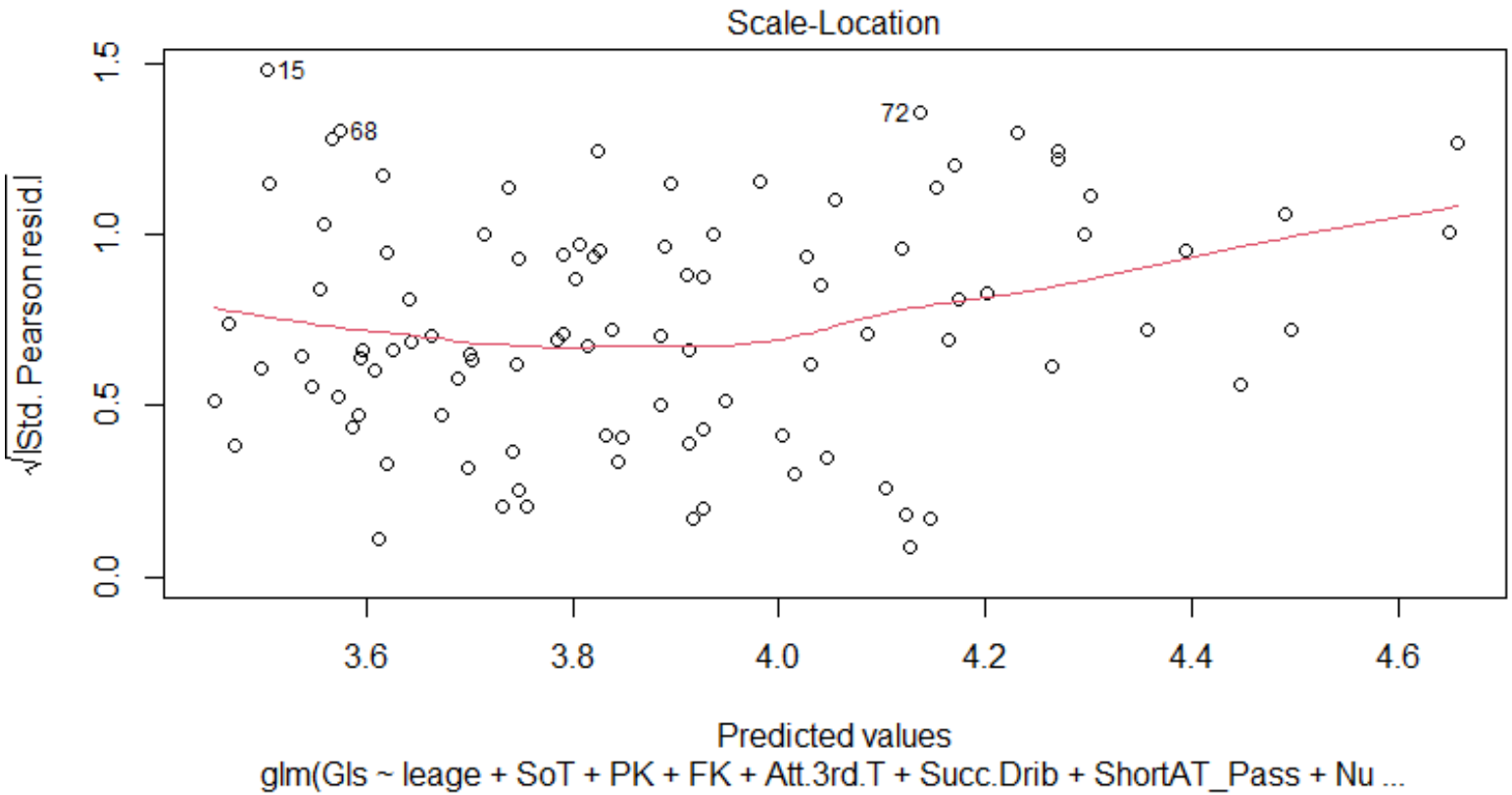

Figure5: Pearson Residual Plots for Poisson regression

Quasi-Poisson regression Output

Table: Quasi-Poisson regression

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.853	0.084	33.861	0
leageBundesliga	0.093	0.043	2.172	0.033
leageLaliga	-0.007	0.039	-0.169	0.866
leageLigue1	0.036	0.039	0.911	0.365
leageSerieA	0.004	0.039	0.109	0.913
SoT	0.005	0.001	7.981	0
PK	0.016	0.006	2.754	0.007
FK	-0.003	0.002	-1.219	0.226
Att.3rd.T	-0.00001	0.00002	-0.348	0.729
Succ.Drib	-0.0005	0.0002	-2.031	0.045
ShortAT_Pass	0.00003	0.00001	2.314	0.023
NumAtt	0.073	0.013	5.791	0.00000

Dispersion parameter= 0.6640285. Residual Deviance= 58.267

Table: Exponent of the estimate sof Q-P output

	Exponent Estimate
(Intercept)	17.343
leageBundesliga	1.097
leageLaliga	0.993
leageLigue1	1.036
leageSerieA	1.004
SoT	1.005
PK	1.016
FK	0.997
Att.3rd.T	1.000
Succ.Drib	1.000
ShortAT_Pass	1.000
NumAtt	1.075

Effect Plot

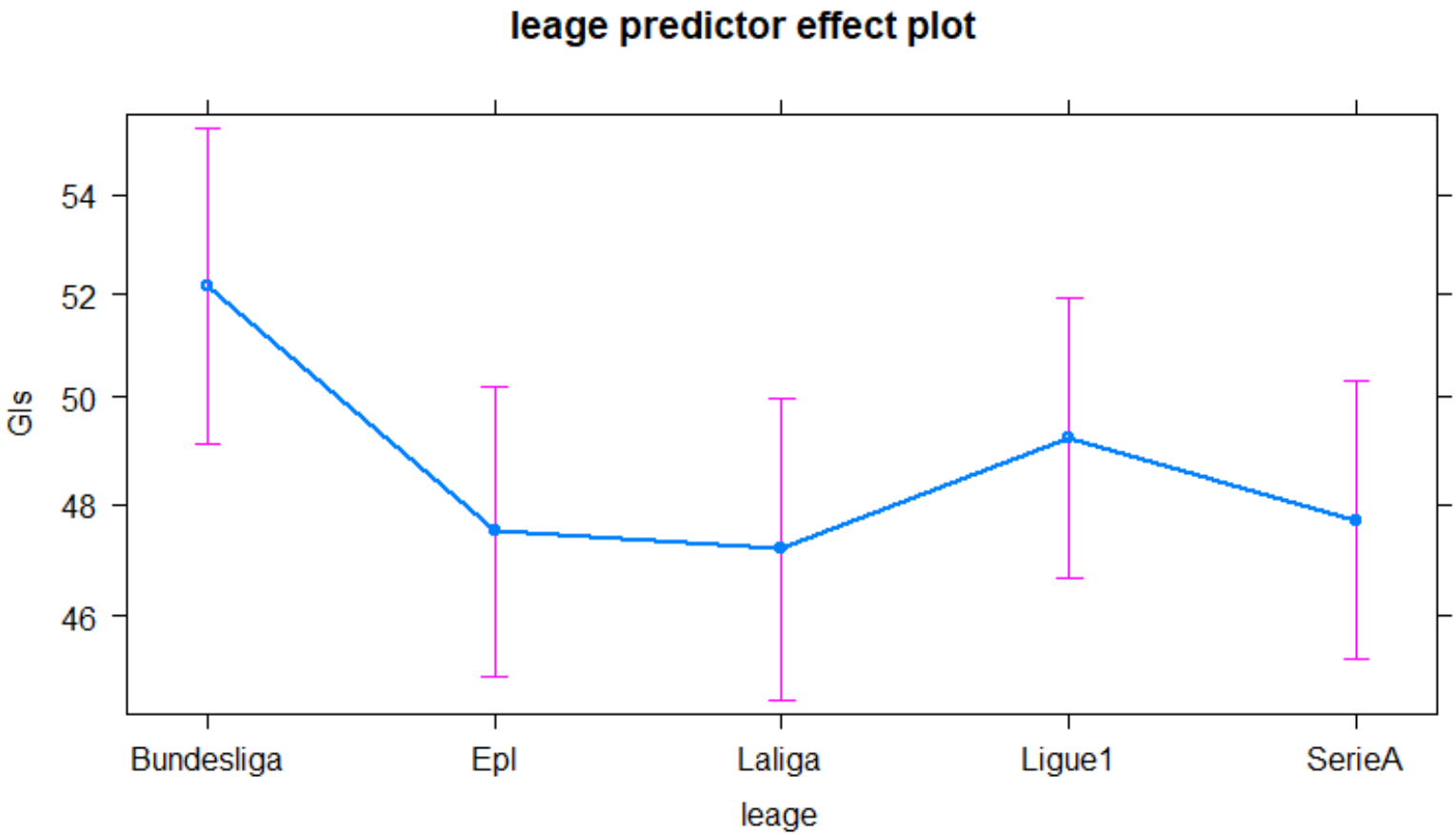

Figure6: League Effect Plot for Poisson regression

Pearson residual plot

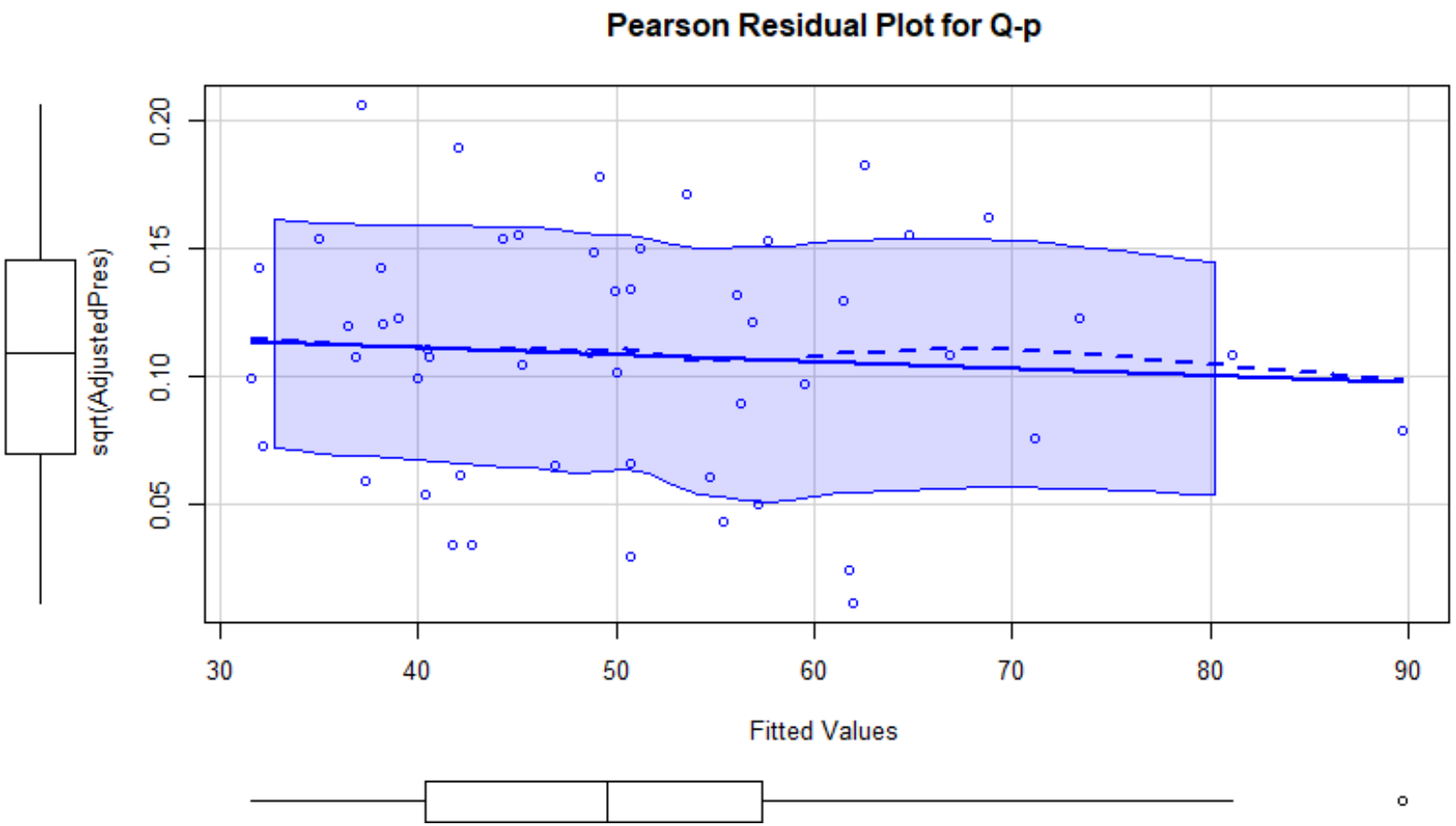

Figure7: Pearson Residual Plots for Q-P regression

Conclusion and Further Studies

Conclusion and Further Studies

- Based on our results we found that there was a significant difference in goals between Bundesliga and other leagues. Among the other leagues there is no significance in goals between them.

- Interestingly, we found that Shots on Target, Penalty Kicks, Successful Dribbles, Short Attacking Passes and Number of Attackers were significant.

- Further studies on this topic would include analysis of multiple seasons rather than just one season. We believe this would give us better results because it would lessen some of the outliers and biases that are present in just one season.

Reference

- Wallace, Jarryd Norton, Kevin. (2013). Evolution of World Cup soccer final games 1966-2010: Game structure, speed and play patterns. Journal of science and medicine in sport / Sports Medicine Australia. 17. 10.1016/j.jsams.2013.03.016.
- `https://fbref.com`

Thank You