International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2016): 79.57 | Impact Factor (2015): 6.391

Prediction in Football Using Poisson Regression Model

A. Yawe[1], H. A. Odiniya[2]

[1]Department of Mathematics and Statistics, Federal University Wukari, P.M.B. 2010, Wukari, Taraba State Nigeria

[2]Information and Communication Technology Department Federal University Wukari, P.M.B. 2010, Wukari, Taraba State Nigeria

Abstract: ***In this work, we will basically use data of some football players in the European Premier League game gotten from the website of fantasy premier league to see if Poisson Regression Model can fit into the data and to come up with some predictions about the selected players. The data used for the analysis in this is work was gotten from the website http://www.fantasypremierleague.com.***

Keywords: Poisson Regression Model, prediction

1. Introduction

Modeling football has become a popular challenge over the recent few years. As such different models have been proposed in order to understand the characteristics that make a team to lose or win in a game or predicting the score of a match. People who watch sports understand its random nature. Therefore can mathematical models be used to predict the outcome of a game? Erlandson *etal, (2016)* predicted football scores of various teams for a league via poisson regression model. The goal here is to look into the possible model that could be used to make predictions in a football and to use the available data we havegot to make inferences about the number of goals each player in a team can score and to possibly infer the match outcome of the team in a match.

2. Poisson Process

A process can be define as an event that evolved over time intending to achieve a goal. In general terms, the time period ranges from 0 to T. Within this time interval events may be happening at different points along the way that may have some impact on the eventual value of the process. A process can either be deterministic or stochastic.

3. Random Variable

Blitzstein and Jessica (2014) defined random variable as a function that maps a sample space S to a real number R in a given trial. Many often denote random variable with a capital letter,say X, though it may not be necessarily compulsory. A random variable maps elements from a domain of a sample space to a range of real line.

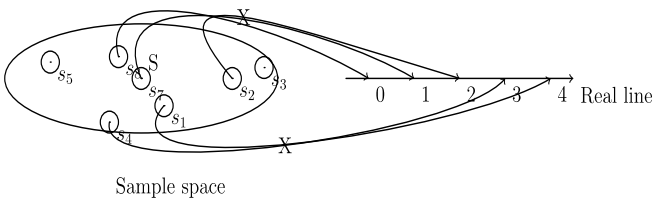

Figure 1: A random variable maps elements from a domain of a sample space to a range of real line.

The figure 1 shows a sample space S containing some elements mapped by a function to a real line with values 0, 1, 2, 3 and 4. The random movement here is defined by the probability function that follows the sample space.

Consequently, the random variable X allocates number values $X(s)$ from the sample S to every result of the trial. The random movement is realized as a result of the probability function P defined on the sample space and at the same time the way the random variable is being assigned. The real number is deterministic as represented in the figure above(Blitzstein and Jessica, 2014). A simple example for this is the example of a coin tosses. Suppose that a fair coin is tossed twice, the sample space for this coin tosses is {HH, TH, HT, TT} where H denotes head and T denotes tail. Now each random combination in the sample space provides a numeric value summary. For instance suppose that the random variable X is the number of heads in the trial. Then X has possible values of 0, 1 and 2. Now seen as a function, X allocates the value 0 to the result TT, 1 to the result of HT and TH, and 2 to the result of HH. That is

$$P\{X = 2\} = \{HH\} = \frac{1}{4}$$
$$P\{X = 1\} = \{TH\} = \{HT\} = \frac{1}{2}$$
$$P\{X = 0\} = \{TT\} = \frac{1}{4}$$

Random variable can either be discrete or continuous. In the example above we could see that the random variable Xtakes on a finite number of possible values 0,1 and 2. This kind of random variable is referred to as discrete. Nevertheless, some random variables exist that does not take finite or countable values, which are known as continuous random variables. But for this work, we shall only discuss about the discrete since the scores in football games are countable.

3.1 Discrete Random Variable

Following Blitzstein and Jessica (2014) a random variable X is discrete if given any countable specific number of values $b_1,b_2,...,b_n$ or countable unspecified number of values $b_1,b_2,...$ such that the probability of X taken the values b_i for some i is the same as unity i.e.$P(X = b_i \; for \; some \; i) = 1$.

Suppose that X is a discrete random variable, then for any value x of either the countably specific number of values or the countably unspecified number of values, such that the probability that the random variable X takes the value of x is greater than zero($P(X = x) > 0$), then x is called a support of X.

Note that: x is an event while X is a random variable. We can only take the probability of an event not a random variable.
A process that can be described by the change of some random variable over time is known as stochastic process. A stochastic process can either be counting or continuous.

3.2 Counting Process

According to Ross (2003) a stochastic process $\{N(t), t \geq 0\}$, where $N(t)$ is the number of specific events by time$t$, is a counting process if the following hold:
(i) $N(t)$ is positive $(N(t) \geq 0)$
(ii) $N(t)$ is integer valued
(iii) If $s < t$then $N(s) \leq N(t)$.
(iv) For $s < t$, $N(t) - N(s) < t$ is the number of events in the time interval $(s, t]$ $0 < s < t < T$

Following Gardner (2011), the goals that are scored by players in the game of a football are random. However, a close observation at the time and how these goals are being scored in the game can be seen as a timeline of events that can be represented in one dimensional scatter plot of time interval t, between 0 to 90 minutes. In a game that for instance three goals were scored in the 20th, 45th and 80th minutes by a player, then each goal can be represented on a 1-dimensional plot by

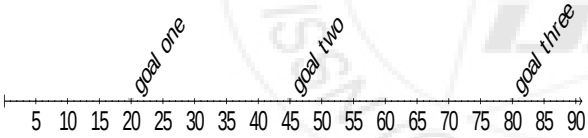

Figure 2:1-dimensional plot for three goals

Now suppose that the total of 6 goals were scored in the game by different players in the 5th, 20th, 30th, 45th, 60th and 80th minutes

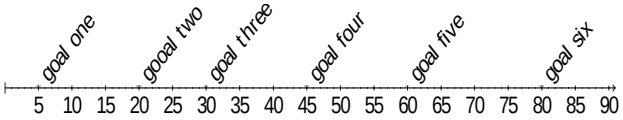

Figure 3:1-dimensional plot for six goals

The selection of these random goals can be viewed as counting process. A counting process has independent increment if the numbers of events happening at different time interval are not dependent. For instance in football game where $N(t)$ is the number of goals by a player at time t can be justifed as independently increasing if we believe that the players chances of scoring at a time does not depend on his performance before that time. "It will not be justified if we believe in 'hot streak' or 'slumps'" (Ross, 2003). A counting process has stationary incrementif the distributions of the number of events that occur in a time depend on the time interval. In football this may hold over a smaller time range since football game is done in a small time range of 90 minutes (Ross, 2003).

A counting process $\{N(t), t \geq 0\}$ is said to be a Poisson process with parameter (λ) where λ > 0 is the rate of the event occurring, if these conditions hold:
(i) $N(t = 0) = 0$
(ii) $N(t)$: $t \geq 0$has an independent increment
(iii) $N(t) - N(s)$is the number of events in the time interval $(s, t]$ $0 < s < t < T$ (Ross, 2003) .

Looking at the counting process illustrated in the timeline in figure 3, it is obvious to see that at time $t = 0$ the number of goals scored $N(t)$ is zero

$$N(t = 0) = 0 \quad (1)$$

at$t$ = 90,

$$N(t = 90) = 6 \quad (2)$$

and suppose that s = 25 and t = 70

$$N(t = 70) = 5 \quad (3)$$

$$N(s = 25) = 2 \quad (4)$$

$N(t) - N(s) = 5 - 2 = 3$ (the number of goals between 25th - 70th minutes)

Hence the goals scored in the intervals between 0 and 90 are seen as Poisson.

4. Poisson Regression Model

Poisson Regression Model is a probability distribution model that is used to model count data. It is a random component of generalized linear model. The model is derived from Poisson distribution by using the same density function and variable that defines how the events happen. Suppose that $y_i$are the scalar outcome response Poisson random variables whose mean depend on the vector of predictor variables X_i. This implies that y_i is condition on X_ii.e. $y_i|X_i$. Then by Poisson distribution, the probability density of y_i condition on X_i is given by

$$f(y_i|X_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \quad (5)$$

$\lambda_i$are the mean vectors of (X_i,β) and the expectation of y_i condition on vector X_i. That is $\lambda = \lambda(X_i, \beta) = E(y_i|X_i)$ where y_i belongs to the set of natural numbers.

$$E(y_i|X_i) = \lambda(X_i) \quad (6)$$

If X_i is not constant, the variance

$$Var(y_i|X_i) = \lambda(X_i) \quad (7)$$

4.1 Log- Linear Models

Suppose that Y_i where $i = 1,2,3, \ldots, n$is a Poisson with mean λ_i and the mean depend on some vector of regressorsX_i. Then λ_i could be written as

$$\lambda_i = X_i'\beta \quad (8)$$

Looking at the above equation we could envisage that the component on right hand side of the equation can be any real value but the Poisson mean on the left hand side is limited to

only positive real value since it represents expected counts. To tackle this kind of problem, we need the logarithm of the mean λ_i with the linear model. That is

$$\log\lambda_i = X_i'\beta \quad (9)$$

Since the log of the expected value of Y is a linear function of explanatory variables and the expected value of Y is multiplicative function of X;

$$\lambda = \exp(\alpha + \beta x) = e^{\alpha}e^{\beta x} \quad (10)$$

5. Poisson Regression for Count Data in R

In this section, we shall use the history data of some football players gotten from Fantasy Premier League Website to apply Poisson regression model. The data for the football players in English Premier League can be found on the internet. For this work, we shall consider only factors namely, play ground (that is either home or away) and the team strength (that is whether good or poor team). We shall use their data collected for at least the first twenty games of the 2014/2015 season. The data is saved in an Excel spreadsheet as "Comma Separated Value (CSV)".

5.1 Home Advantage Parameter

The home advantage parameter would give the home player a weighting due to the fact that in football it is believed that whoever is playing at home in a game has advantage over their opponent. It could be as a result of having larger number offans at home ground or because they are used to playing at the ground (Gardner, 2011).

5.2 Team Strength Parameter

The strength of the team played against could determine the player's performance. This is because a player tend to display better against a poor team than a good team due to the fact that the counter attack experienced from poor teams are less compared to the counter attack by good teams. Thus we would give an extra weight to the player playing against a poor team.

5.3 Players

In this work we shall be using an eleven player team, comprising of a goal keeper, three defenders, five midfielders, and two strikers. The players are selected from different clubs in the English Premier League. For the purpose of this work, each player is been assigned an identity number. Table 1 below shows a summary of the players, their identities, clubs and playing position.

The data to be used in the work is saved in a CSV file format on an Excel spread- sheet. The data file contains the name and identity of each player, the goals score for each game week by the player, the place of the match (whether home or away) and the strength of the team played against among others for twenty game week. This spreadsheet can be imported using the command below:

Table 1: Selected players

S/N	Player	Identity	Club	Position
1	Diego Costa	1	Chelsea	Striker
2	Wayne Rooney	2	Manchester United	Striker
3	David Silva	3	Manchester City	Midfielder
4	Stewart Downing	4	Liverpool	Midfielder
5	GrazianoPelle	5	Southampton	Midfielder
6	Alexander Song	6	West Ham	Midfielder
7	Raheem Sterling	7	Liverpool	Midfielder
8	John Terry	8	Chelsea	Defender
9	Jose Fonte	9	Southampton	Defender
10	Phil Jones	10	Manchester United	Defender
11	David DeGea	11	Manchester United	Goal Keeper

```
>mydata1 <- read.csv(file="teamdata.csv")
> mydata1
              name id score  t team.against place point strength
1    Diego Coasta  1     1  1   BUR(A) 3-1     0     6        1
2    Diego Coasta  1     1  2   LEI(H) 2-0     1     6        1
...
```

The table contains the data history of each of the eleven players for twenty-week games in the 2014/2015 Premier League Season. The data is in the form of a longitudinal data. Dummy variable is used to indicate whether a match was played on the home ground or away and also, whether a rival team is a good or poor team.

5.4 Estimating the Parameter Vector β

To estimate the parameters using the football data inputted into R, we use the generalized linear model function called 'glm'. The family of the glm is indicated as Poisson with link log. However, in this work, we shall use the quasi-Poisson in other to avoid over-dispersion.

```
>model1 <- glm(score ~ place + strength, data=mydata1, family=quasipoisson(link=
>summary(model1)
```

Table 2: Estimates for pameters β

Coefficients	Estimates
Intercepts	-2.2737
Place	0.3445
Strength	0.6466

We can get the estimates for the within effects for each player by using the code below:

```
>model <- glm(score ~ place + strength + factor(id)-1, data=mydata1, family=quas
>summary(model)
```

5.5 Prediction

We would predict four different scenarios to see what the model has to say about the data. The first scenario playing at home ground with a poor team, second is playing at an away ground with a good team, third is playing at a home ground with a good team and fourth is playing on an away ground with a poor team. The predicted plots for the four cases outlined are shown below:

Table 3: Estimates for Individual Within Effect

Coefficients	Estimates
place	0.3946
strength	0.5510
1	-0.7837
2	-1.4658
3	-1.5491
4	-2.1380

5	-1.7140
6	-20.8726
7	-2.2265
8	-2.9692
9	-20.9053
10	-20.9259
11	-20.9564

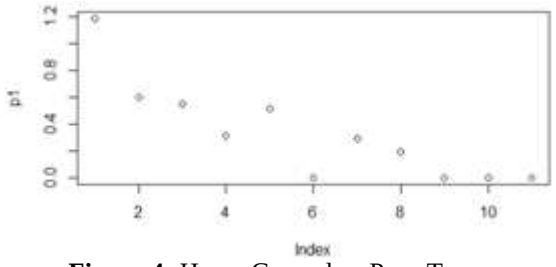

Figure 4: Home Ground vs Poor Team

Figure 4: This shows the predictions for each player as they would play on their home ground with a poor team. It is evident that the team has high chances of scoring more goals

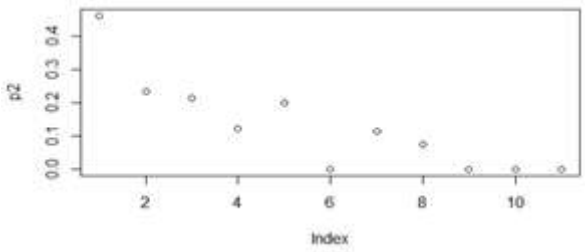

Figure 5: Away Ground vs Good Team

Figure 5: Shows the predictions for each player as they would play outside their home and against a strong team. The plot shows that the team may hardly record goals.

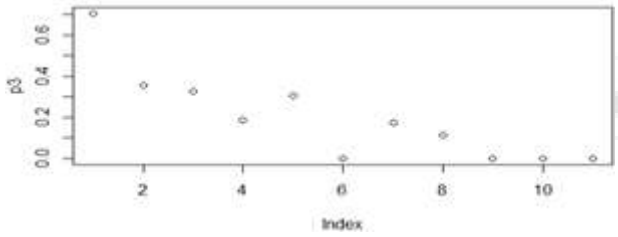

Figure 6: Home Ground vs Good Team

Figure 6: Shows the predictions for each player as they would play at home with a good team. It is evident that the team could score slightly.

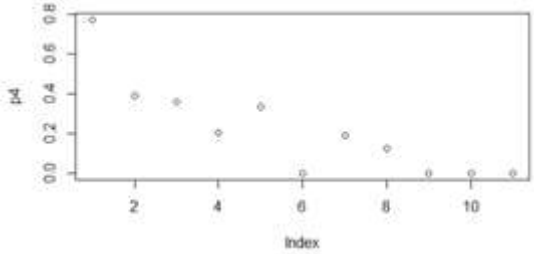

Figure 7: Away Ground vs Poor Team

Figure 7: This shows the predictions for each player as they would play on away ground with a poor team. The plots show that the team could record more goals than in figure 3
From the plots above, we could see that players with identity 1, 2, 3, 4, 5 and 7 are comprised of strikers and midfielders. This shows that they have edge of scoring goals over players with identity 6, 8, 9, 10 and 11which are comprised of defenders and a goal Keeper. This shows that Poisson regression model fits into football data.

6. Conclusion

The result in this work shows that the randomness in football can be studied using Poisson regression model. Though there could be over estimation of goals in the predictions made in this work due to the fact that other factors where not considered; like the defense and attack force of the opposition team.

References

[1] Blitzstein, K. J. and Jessica, H. "Introduction to probability". CRC Press, Taylor and Francis Group 6000 vroken Sound Parkway NW, Suite 300 Boca Raton, FL 33487-2742. 2014
[2] Erlandson, F. S., Adriano, K. S., Ciro, A. O. F. and Francisco L. "Predicting football scores via Poisson regression model: application to the national football league".commun. Stat. Appl. Methods 23:297-319. 2016.
[3] Gardner, A. J.."Modelling and Simulating Football Results kernel description". 2011
[4] Rodriguez, G. "Poisson model for count data". Accessed: 2015-03-20. 2007
[5] Ross, M. S."Introduction to probabilty models" Academy Press, 525 B Street, Suite 1900, San Diego, CA 92101-4495, USA. 2003.

Author Profile

A.Yaweholds a Master of Science degree in Statistics and Operations Research from University of Essex, UK in 2015 and Bachelor of Technology degree in Mathematics from Modibbo Adama University of Technology Yola, Nigeria in 2012. Currently, a Lecturer, Department of Mathematics and Statistics, Federal University Wukari, Nigeria.

H. A. Odiniya Received the M.Sc. and B.Sc degrees in Operation Research from ModibboAdama University of Technology Yola in 2017 and 2011 respectively and is currently working as a System Analyst at Information and Communication Technology Department Federal University Wukari, Nigeria.