SORT 43 (1) January-June 2019, 95-112 DOI: 10.2436/20.8080.02.81

A probabilistic model for explaining the points achieved by a team in football competition. Forecasting and regression with applications to the Spanish league

Emilio Gómez-Déniz[1], Nancy Dávila-Cárdenes[1]
and José María Pérez-Sánchez[2]

Abstract

In the last decades, a lot of research papers applying statistical methods for analysing sports data have been published. Football, also called soccer, is one of the most popular sports all over the world organised in national championships in a round robin format in which the team reaching the most points at the end of the tournament wins the competition. The aim of this work is to develop a suitable probability model for studying the points achieved by a team in a football match. For this purpose, we built a discrete probability distribution taking values, zero for losing, one for a draw and three for a victory. We test its performance using data from the Spanish Football League (First division) during the 2013-14 season. Furthermore, the model provides an attractive framework for predicting points and incorporating covariates in order to study the factors affecting the points achieved by the teams.

MSC: 62J02, 62J20, 62F15.

Keywords: Covariate, football data, forecasting, regression, sport statistics, truncated distribution, weighted distribution.

1. Introduction

Football or soccer sparks interest not only among its supporters or fans, but has also become one of the most profitable industries, with a significant economic impact in infrastructure development, TV rights, sponsorships and transfers of players. According to Marian Otamendi, director of the World Football Summit, the international event of the football industry, gathering the most influential professionals, football could become the 17th largest world economy. Beyond the game, the growth in revenues and

[1] Department of Quantitative Methods and T*i*DES Institute. University of Las Palmas de Gran Canaria, Spain.
[2] Department of Applied Economic Analysis. University of Las Palmas de Gran Canaria, Spain.
Received: July 2018
Accepted: December 2018

the worldwide interest in football prove a successful and lucrative industry in which the aggregate revenue for the top 20 Money League clubs rose 6 percent to 7.9 billion € in 2016/17 (from Deloitte Football Money League 2018). Therefore, football and money go hand in hand and it is also interesting to see, as a simple example, how the emergence of the Chinese Super League and its financially and politically powerful clubs impact on the established European football business order.

Focusing on the game itself, a football competition is played under two basic types of tournaments around the world, the round robin and the knock-out. In the first one, each team plays against each opponent twice in home and away. The possible outcomes are win, draw or loss and the teams receive three, one or none points respectively depending on the result. At the end of a season, the team with the largest number of points wins the championship. This sport has become a multi-billion dollar business, where tactics are basic to the game and with many styles and playing formations available (see for example, (Brillinger, 2008)). Statisticians started to create models to analyse the several aspects involved in a football match, from predicting the outcome of soccer games to determine the best playing strategies, see Díaz and Núñez (2010) and Louzada, Suzuki and Salasar (2014), among others. According to Karlis and Ntzoufras (2000), research in soccer statistics can be divided into three main categories. The first one models the outcome of a game what can be used for ranking soccer teams and it may be extended to quantify the home effect. The second one investigates models for predicting about the number of goals scored by each team, and the third one concentrates in modelling other characteristics of the game. As pointed out by Rue and Salvesen (2006), the outcome of a soccer match depends on many factors, among these are: the home-away ground effect, the effect of injured players, psychological effects, etc. A good knowledge about these factors only determines the result up to significant, but not too dominant, random components. Other papers have focused on modelling football outcomes through the number of goals scored as Karlis and Ntzoufras (2003) who made use of the correlation of the goals scored by the two teams. Also, Rue and Salvesen (2006) predicted the outcome using a Bayesian methodology, whose predictive accuracy is better than other techniques. Finally, Baio and Blangiado (2010) predicted football results making use of a Bayesian hierarchical model.

The method for assigning three points for a win, no points awarded to the losing team and one point assigned to each team if the game ends with a draw, is a standard scoring system used in many sports leagues and tournaments, especially in football, field hockey, the rugby union, ice hockey, among others. However, the scoring system has changed over time. Many leagues and competitions originally awarded two points for a win and one point for a draw, before switching to the three points for a win system. The increase in rewards for a win from two to three points was adopted in 1995. Hon and Parinduri (2016), using regression discontinuity design as the empirical strategy, did not find evidence that the three-point rule makes games more decisive, increases the number of goals, or decreases goal differences, they found some evidence that the three-point rule increases the second-half goals of the losing first-half team. In this paper, far from pre-

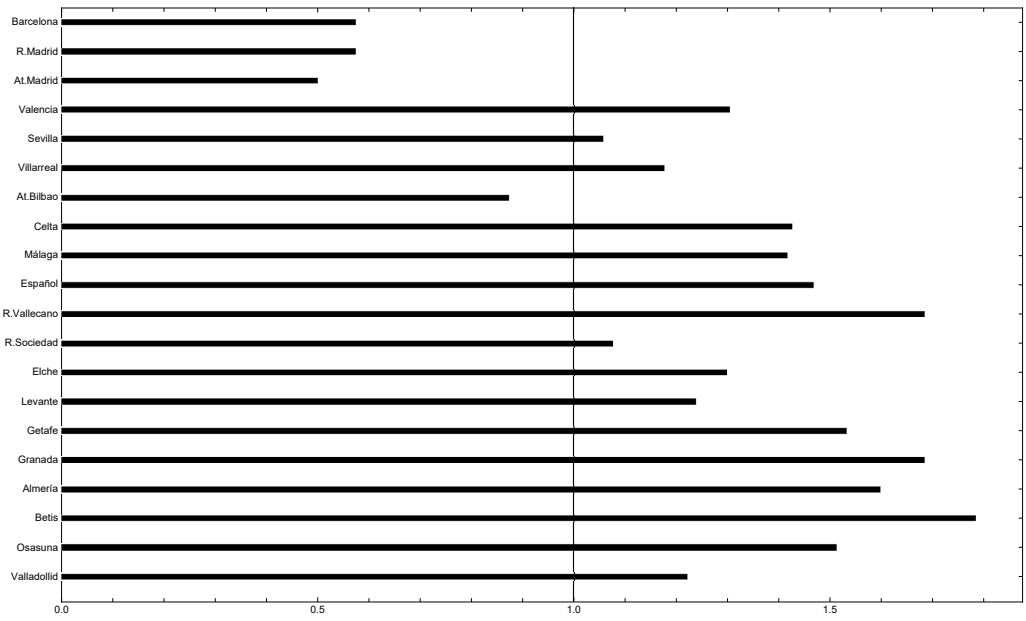

Figure 1: *Index of dispersion for the different teams in the Spanish League.*

dicting football results or analysing the effect of the scoring system in the results, we consider soccer matches played in a league in which the teams play against each other twice (home and away) where many explanatory variables may influence the result of a forthcoming soccer match. In this context, we analyse the factors that could affect the points achieved by a football team using data from the Spanish League during the 2013-2014 season. A similar analysis could be done to other leagues and sports for which a data base were available.

Empirical analysis shows that the sequence of points for teams with a lot of points, and therefore fighting for the title of the competition, is characterized for being under-dispersed (variance lower than the mean) while the teams with less points at the end of the competition show over-dispersion (variance larger than the mean). Let X be the random variable which gives us the sequence of points achieved by a team in a competition. The index of dispersion is defined as $\mathrm{ID} = \mathrm{var}(X)/\mathrm{E}(X)$. This value is represented in Figure 1 for the twenty teams of the Spanish Football League at the end of the competition in 2013-14 season. As we can see, this index is lower than 1 for the five best teams of the competition while is larger than 1 for the worst teams.

In this work, we present a probability model to analyse the points achieved by a team in a football match competition. That is, we propose a probability model which takes values in the set $\{0,1,2,3\}$ and with zero mass probability in $x = 2$. Furthermore, the model accommodates for over-dispersion and under-dispersion and it is suitable for incorporating covariates. The proposed model is simple and the estimation of the

parameters is easily obtained. Therefore, it is a candidate for fitting data sets of points in football match competitions.

The rest of this paper is organised as follows. The main model together with some of its most important properties are developed in Section 2. In Section 3, an application to the Spanish Football League in 2013-14 season is given. Finally, summary and discussion of the results are shown in the last Section.

2. Probabilistic model for points

The Poisson distribution represents a simple model as a starting point to construct a probability mass function (pmf) in the scenario we are considering. In football sport the number of goals scored by each team in a match has been assumed to follow a Poisson distribution by numerous authors. Some examples in which the Poisson distribution has been used to predict football results are Karlis and Ntzoufras (2000), Greenhough et al. (2002) and Saraivaa et al. (2016). However, to our knowledge, the distribution of the number of points, which is a discrete variable, has not been formally treated. Let us to start with the classical Poisson distribution whose pmf is given by,

$$f_\theta(x) = \frac{\theta^x \exp(-\theta)}{x!}, \quad x = 0, 1, \ldots, \ \theta > 0. \tag{1}$$

We need a random variable X which takes only 4 values, to say 0,1,2 and 3 and with the constraint that for $x = 2$ the mass of probability should be zero. Therefore, it has a two-parameter pmf of the form: $P(X = 0) = 1 - p - q$, $P(X = 1) = q$, $P(X = 2) = 0$ and $P(X = 3) = p$, with $0 < p < 1$, $0 < q < 1$, $p + q < 1$. The "probability generating function (pgf)" is $g(t) = 1 - p - q + qt + pt^3$ and $E(X) = q + 3p$. In order to simplify the model, we attend to the particular case $p = \theta^3\kappa(\theta)$ and $q = \theta\kappa(\theta)$, where

$$\kappa(\theta) = \frac{6}{24 + \theta(6 + \theta^2)}. \tag{2}$$

Thus, the expression

$$g_\theta(x) = \kappa(\theta)(x-2)^2\frac{\theta^x}{x!}, \quad x = 0, 1, 3 \tag{3}$$

defines a genuine pmf with support in $\mathcal{X} = \{0, 1, 3\}$. Recall that for a distribution with pmf $f_\theta(x)$, X with support in $\mathcal{X}$, depending on a vector of parameteres $\theta \in \Theta$, we can construct a new distribution with pmf (see for instance Fisher, 1934, Patil and Rao, 1978 and Harandi and Alamtsaz, 2013) using a weighted function, $w(x) > 0$,

$$g_\theta(x) = \frac{w(x)}{E_{f_\theta(X)}(w(X))} f_\theta(x), \tag{4}$$

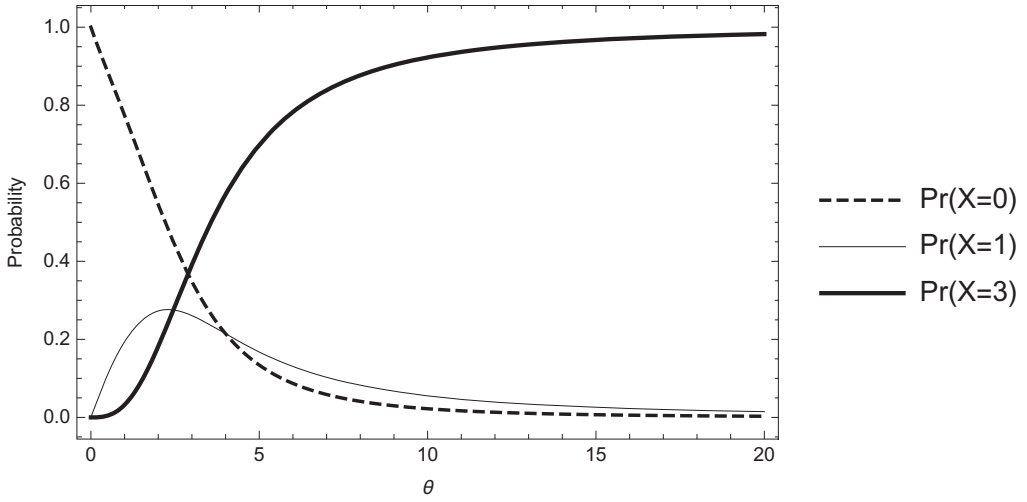

Figure 2: *Probabilities of victory, draw and defeat depending on θ.*

where it is assumed that $E_{f_\theta(X)}(w(X)) < \infty$, and w is a weighted function depending on X. Now, it is easy to see that the pmf given in (3) is a weighted version of the pmf given in (1) by taking $w(x) = (x-2)^2$ and restricting (truncating) its support to take only the values 0, 1 and 3. Therefore we are using truncation and weighting, where the latter can be viewed as a particular case of the first (see Johnson, Kemp and Kotz, 2005, p. 63 for details).

Figure 2 shows the graph of the values of p, q and $1-p-q$, i.e. the values of the probability of victory, draw and defeat, in the definition domain of θ parameter.

It can be easily proved that the following chains of inequalities are satisfied among the probabilities of the three events that are intended to be modelled through the pmf given in (3):

$$\begin{aligned}
&0 < \theta < 2.45: && \Pr(X=0) > \Pr(X=1) > \Pr(X=3),\\
&2.45 < \theta < 2.88: && \Pr(X=0) > \Pr(X=3) > \Pr(X=1),\\
&2.88 < \theta < 4: && \Pr(X=3) > \Pr(X=0) > \Pr(X=1),\\
&\theta > 4: && \Pr(X=3) > \Pr(X=1) > \Pr(X=0).
\end{aligned}$$

Therefore, the teams that have the expectation of playing Champions League will be characterized by the achievement of points that make the θ parameter greater than 4. In contrast, teams that only aspire to maintain the category are characterized by θ values less than 2.45. Hence, the θ parameter can be interpreted as the value that will position a team in the four areas in which a football competition can be divided: Champions, Euroleague, no-relegation and relegation zones.

2.1. Statistical properties

The moments can be obtained from the pgf. In particular, the mean and second order moment about zero are given by

$$E(X) = \frac{\kappa(\theta)}{2}\theta(2+\theta^2), \tag{5}$$

$$E(X^2) = \frac{\kappa(\theta)}{2}\theta(2+3\theta^2). \tag{6}$$

Using (5) and (6) we get the variance, given by

$$var(X) = \frac{\kappa(\theta)^2}{3}2\theta\left[6+\theta^2(9+\theta)\right] \tag{7}$$

and some computations provide the index of dispersion, which is

$$ID = \frac{var(X)}{E(X)} = \frac{4\kappa(\theta)\left[6+\theta^2(9+\theta)\right]}{3(2+\theta^2)}. \tag{8}$$

Figure 3 shows the ID given in (8) for some values of the support of the parameter θ.

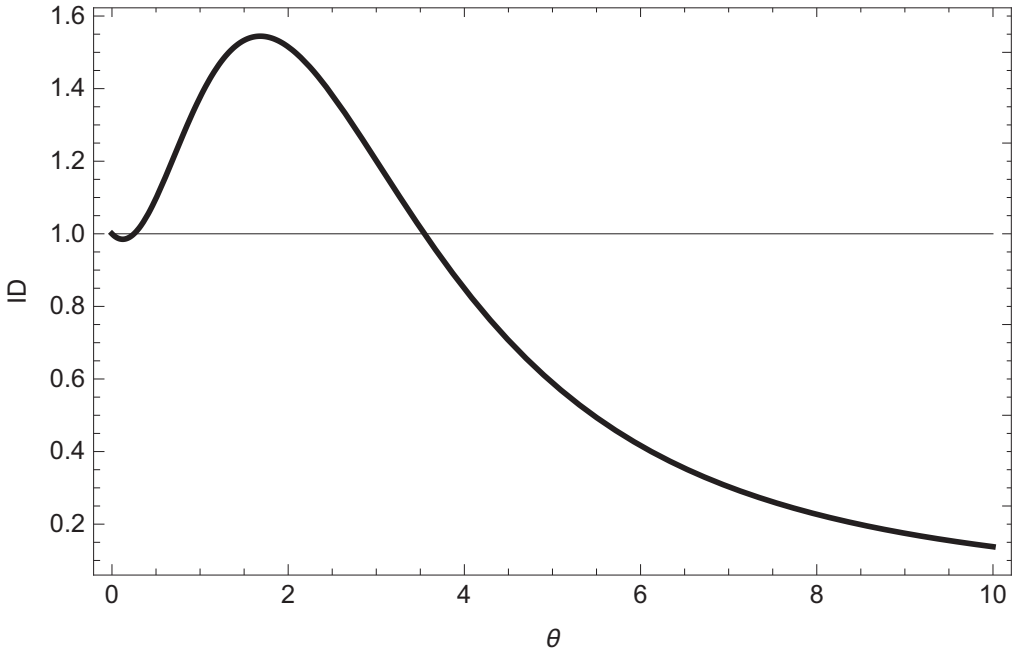

Figure 3: *Index of dispersion of the probability model depending on θ.*

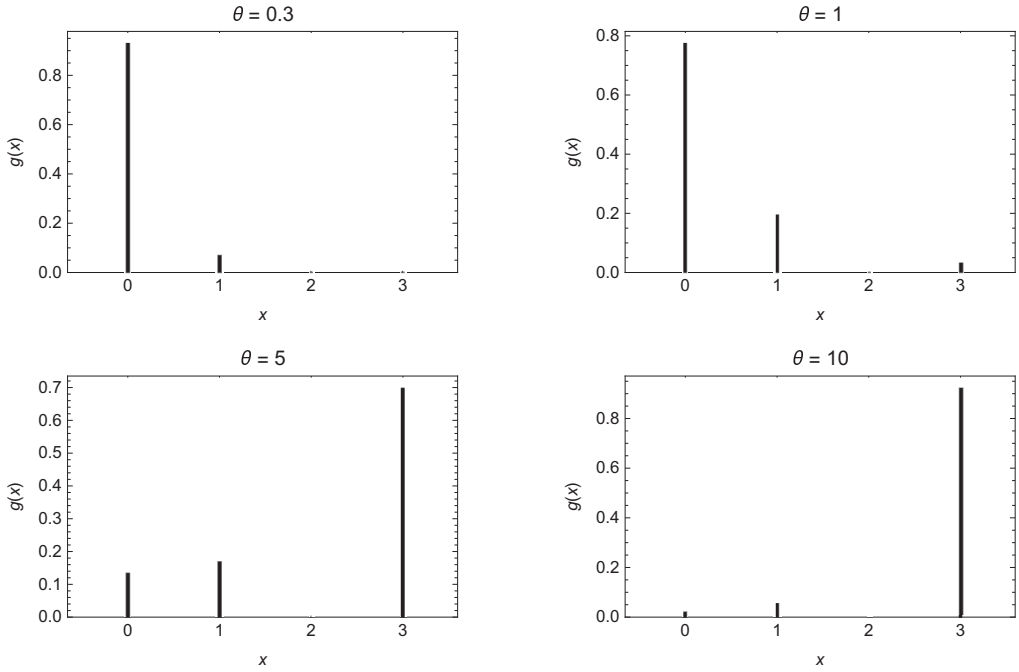

Figure 4: *Plot of the pmf in (3) for selected values of the parameter θ.*

It is simple to verify that the pmf accommodates over-dispersion (variance larger than the mean) when $0.250081 < \theta < 3.54676$ and under-dispersion when $0 < \theta < 0.250081$ and when $\theta > 3.54676$. Furthermore, the maximum value of the ID is reached for $\theta = 1.68365$, taking the value of 1.544.

Furthermore, the relation connecting the cumulants $k_{[r]}$ and the moments about the origin μ_r can be obtained using expression (8) in Noack (1950). Relations between factorial-cumulants and cumulants can also be given using results in Khatri (1959). See also Johnson et al. (2005, p. 77).

Figure 4 shows the probability mass function of the proposed model for different parameter values.

In view of this figure, a large value of the parameter θ gives more mass of probability to the value $x = 3$, and then to the victory, which is consistent with the values that give rise to an index of dispersion lower than one. For the teams that finally remain in the leaderboard in the highest positions, the dispersion index is lower than one, as shown in Figure 1. Additionally, for values of the parameter between 0.25 and 3.5 (such as $\theta = 0.3$ and $\theta = 1$) the distribution assigns more weight to the values $x = 0$ and $x = 1$, therefore it represents teams which have obtained few victories and, in consequence, these teams will remain in the lagging positions at the end of the season.

The cumulative distribution function can be written in terms of the exponential integral function given by $E_n(z) = \int_1^\infty \exp(-zt)/t^n\, dt$, and it results

$$\Pr(X \leq x) = \frac{\kappa(\theta)}{x!}\left[\theta^{x+1}(3-x-\theta+(4+\theta(\theta-3))\exp(\theta)E_{\theta}(-x)\right].$$

2.2. Parameter estimates

In this subsection, two estimation methods for estimating the parameter of the distribution are analysed. First, the method of moments for which let $\tilde{x} = (x_1, x_2, \ldots, x_n)$ be a random sample obtained from model (3). Then, using (5) it is simple to see that the estimator of θ is the real solution of the equation

$$\theta^3(3-\bar{x}) + 6\theta(1-\bar{x}) - 24\bar{x} = 0, \tag{9}$$

where $\bar{x}$ is the sample mean.

Second, the maximum likelihood estimation that it will be used here and where the θ estimator is easy to derive. The log-likelihood function is proportional to

$$\ell(\tilde{x};\theta) \propto n\log\kappa(\theta) + n\bar{x}\log\theta. \tag{10}$$

The likelihood equation obtained from (10) results

$$\theta\kappa'(\theta) + \bar{x}\kappa(\theta) = 0,$$

and provides the unique maximum likelihood estimator of θ, which is the same solution of the equation given in (9). Thus, the moment and the maximum likelihood estimators of the parameter θ are the same.

A little algebra provides the Fisher's information matrix, given by

$$\mathcal{I}(\widehat{\theta}) = E\left[-\frac{d^2\ell(\theta;\tilde{x})}{d\theta^2}\right]_{\theta=\widehat{\theta}} = \frac{2n\kappa(\widehat{\theta})^2}{3\widehat{\theta}}\left[6+\widehat{\theta}^2(9+\widehat{\theta})\right],$$

where $\text{var}(\widehat{\theta}) = \left[\mathcal{I}(\widehat{\theta})\right]^{-1/2}$. The discrete distribution proposed in this work satisfies the regularity conditions (see Lehmann and Casella, 1998, p. 449) under which the unique maximum likelihood estimator $\hat{\theta}$ of θ is consistent and asymptotically normal. They are simply verified in the following way. Firstly, the parameter space $\{0,1,3\}$ is a subset of the real line and the range of x is independent of θ. Additionally, the parameter θ is identifiable, that is, if $\theta_1 \neq \theta_2$ then $\exists x \in \mathcal{X}$ such that $g_{\theta_1}(x) \neq g_{\theta_2}(x)$. By using expression (10) it is easy to show that $E(\frac{\partial \log g_{\theta}(x)}{\partial\theta}) = 0$. Now, because, $\left.\frac{\partial^2\ell(\tilde{x};\theta)}{\partial\theta^2}\right|_{\theta=\widehat{\theta}} < 0$, the Fisher's information is positive. Finally, by taking $M(x) = \frac{\partial^3 \log g_{\theta}(x)}{\partial\theta^3} + 1$, a function which may depend on θ, we have that $\left|\frac{\partial^3 \log g_{\theta}(x)}{\partial\theta^3}\right| \leq M(x)$ and $E(M(x))$ is finite. Therefore the maximum likelihood estimator $\widehat{\theta}$ of θ is consistent and asymptotically normal and

$$\sqrt{n}(\widehat{\theta}-\theta) \xrightarrow{d} N(0,\mathcal{I}^{-1}(\widehat{\theta})),$$

where $N(\cdot,\cdot)$ represents the normal distribution. For details about this assert, the reader can consult Corollary 3.11 in Lehmann and Casella (1998). Due to this, we conclude that the maximum likelihood estimator of θ is asymptotically efficient.

2.3. Including covariates

In this section, we investigate the covariates that may affect the number of points achieved by the teams playing in home (and later away). Let X be a response variable, and let $\boldsymbol{y}$ be an associated $k\times 1$ vector of covariates. For the sake of convenience, we rewrite (3) in another form, so that covariates may be introduced into the model. By equating (5) to μ we get the same equation as the one given in (9). Now, by using Cardano's method of solution of the cubic polynomial equation we get $\theta=\sum_{i=1}^{2} R_i$, where

$$R_i=\sqrt[3]{(-1)^i\sqrt{4\left(\frac{\mu}{\mu-3}\right)^2+\left[\frac{1-\mu}{3(3-\mu)}\right]^3}-\frac{2\mu}{\mu-3}}.$$

The solution for the $\theta\equiv\theta(\mu)$ parameter given above can also be written in another way.[1] A common specification for the mean parameter μ is in terms of exponential functions, ensuring the non-negativity of this parameter. That is,

$$\mu_i=\frac{3\exp\left(\boldsymbol{\beta}^\top\boldsymbol{y}\right)}{1+\exp\left(\boldsymbol{\beta}^\top\boldsymbol{y}\right)}, \tag{12}$$

where $\boldsymbol{y}$ is the vector of covariates and $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_q)^\top$ is an unknown vector of regression coefficients. Expression (12) ensures that the mean is a positive-valued function with support in $[0,3]$. Now, (3) is written as

$$g_{\theta_i}(x_i)=\kappa(\theta(\mu_i))(x_i-2)^2\frac{\theta(\mu_i)^{x_i}}{y_i!},\quad i=1,2,\ldots,n,$$

1. The closed expression for the θ parameter results

$$\theta\equiv\theta(\mu)=\frac{\sqrt[3]{2}\left[\sqrt[3]{2}(3+\mu(\mu-4))-\psi(\mu)^{2/3}\right]}{(\mu-3)\psi(\mu_i)^{1/3}},$$

where

$$\psi(\mu)=6\mu(\mu-3)^2+\sqrt{2(\mu-3)^3(\mu(3+19\mu(\mu-3))-1)}. \tag{11}$$

where again $\kappa(\theta(\mu_i))$ is as in (2). The log-likelihood of the model with covariates is proportional to

$$\ell(\tilde{x};\boldsymbol{\beta}) \propto \sum_{i=1}^{n} \left[\log(\kappa(\theta(\mu_i))) + x_i \log\theta(\mu_i)\right].$$

The normal equations which provide the maximum likelihood estimates of the parameters β_j, $j = 1,\ldots,q$, are

$$\frac{\partial \ell(\tilde{x};\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{1}{\kappa(\theta(\mu_{ij}))} \frac{\partial}{\partial \beta_j} \kappa(\theta(\mu_{ij})) + \sum_{i=1}^{n} \frac{x_{ij}}{\theta(\mu_{ij})} \frac{\partial}{\partial \beta_j} \theta(\mu_{ij}) = 0, \tag{13}$$

for $j = 1,2,\ldots,q$. The second partial derivatives can be seen in the Appendix Section.

Maximising the log-likelihood function (13) with respect to β_j $(j = 1,\ldots,q)$ is simple via the scoring algorithm or Newton-Raphson iteration. The solutions of the nonlinear equations shown in the Appendix provide the maximum likelihood estimates of these parameters. However, these equations cannot be explicitly solved and the solutions may be obtained either by maximising the log-likelihood function or by numerical methods. Different initial values of the parametric space can be considered as a seed points. In this study, the `FindMaximum` function of Mathematica software package v.11.0 (see for instance, Wolfram, 2003 and Ruskeepaa, 2009) was used, although the same results can be obtained by other methods, such as Newton, PrincipalAxis or QuasiNewton (all of which are available in this package), or by other packages such as R, Matlab or WinRats. Finally, the standard errors of the parameter estimates were obtained by inverting the Hessian matrix.

2.4. Marginal effects

The marginal effect reflects the variation of the conditional mean of X due to a one-unit change in the j-th covariate, and is calculated as

$$\frac{\partial \mu_i}{\partial \beta_j} = y_j \mu_i \left(1 - \frac{\mu_i}{3}\right),$$

for $i = 1,\ldots,n$ and $j = 1,\ldots,q$. Thus, the marginal effect indicates that a one-unit change in the j-th regressor increases or decreases the expectation of the points, pointing out that it depends on the sign, positive or negative, of the regressor for each mean. For indicator variables such as y_k, which takes only the value 0 or 1, the marginal effect in term of the odds-ratio is $\exp(\beta_j)$. Therefore, the conditional mean is $\exp(\beta_j)$ times larger if the indicator variable is one rather than zero.

3. Numerical application

In this section, we consider the data corresponding to the points obtained by the 20 teams participating in the First Division of the Spanish Football League in the 2013-14 season. This section is divided into two parts. First, we analyse the predictive capacity of the proposed model without including covariates. So, we study the expected points of the whole season and the expected points and positions based on the first 19 matches of the season (the middle of the competition). Second, we try to identify the significant factors which can explain the expected number of points of the home team by including covariates in the analysis.

3.1. Number of points without covariates

1. Prediction of the final points for the home teams. Figure 5 shows the observed and fitted accumulated points for home teams. The estimated value of the θ parameter is 3.370 and the standard error is 0.131.

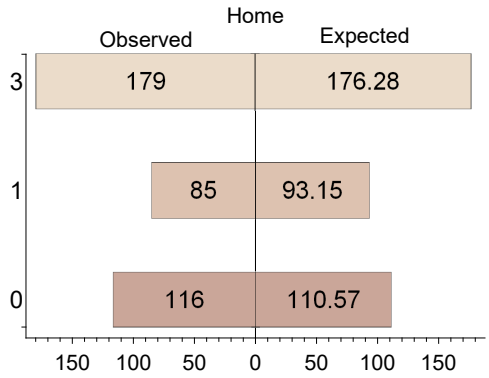

Figure 5: *Observed (left) and fitted (right) home points.*

2. Prediction of the final points based on the first 19 match-days (190 matches). Figure 6 shows the accumulated observed and expected points based on the first 19 match-days.

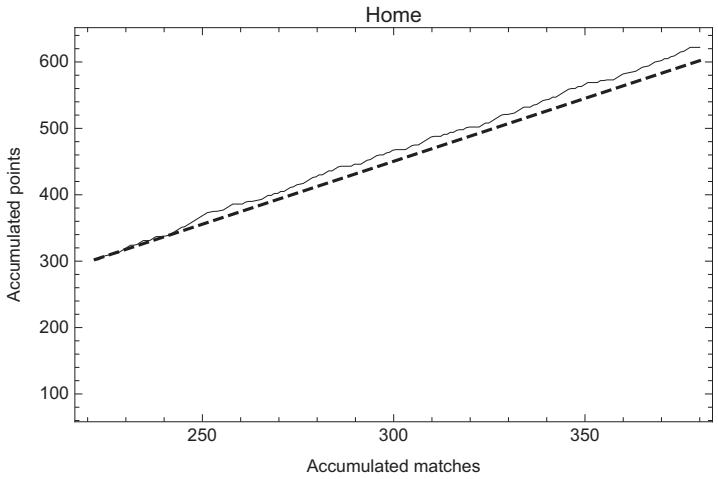

Figure 6: *Accumulated observed points (thin line) and expected (dashed line) based on the first 190 matches given the pmf (3) in the Spanish League.*

3. Prediction of the position of the teams at the end of the competition based on the first 19 match-days of the competition. Table 1 shows the estimated value of θ, the standard error (SE), the value of the maximum of the log-likelihood and the estimated μ parameter. Figure 7 illustrates the prediction of the positions of the teams at the end of the competition. The maximum likelihood estimated value of the θ parameter and the index of dispersion appears between parenthesis.

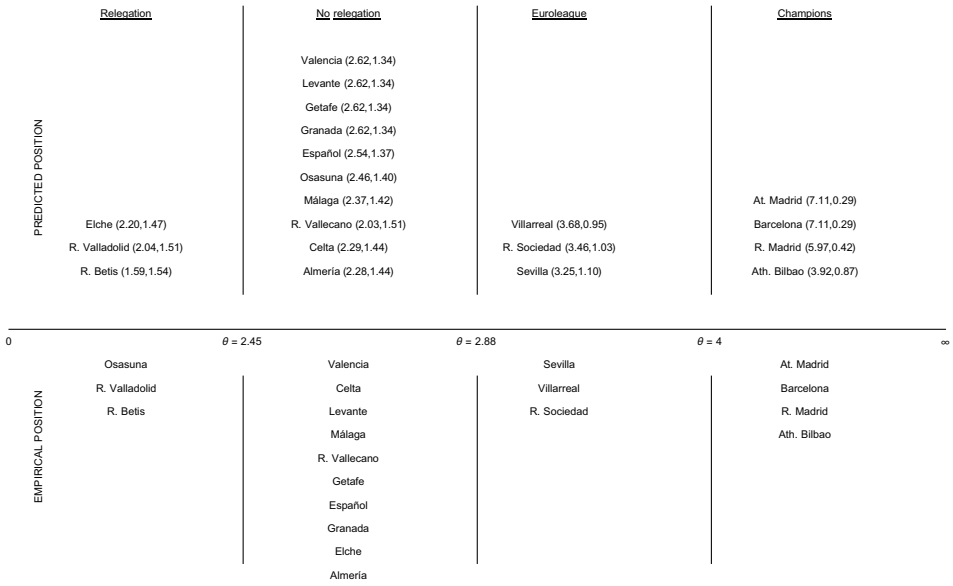

Figure 7: *Prediction based on the first 19 match-days and real final position of the teams at the end of the competition.*

Table 1: *Estimations based on first 19 match-days.*

$\widehat{\theta}$	S.E.	$\ell_{\max}$	$\widehat{\mu}$
3.269	0.179	−199.706	1.584

3.2. Number of points including covariates

First, we briefly describe all the variables that have been considered in the study in order to analyse the factors involved in the total points achieved by a team in a competition, we consider four groups of variables. Those related to the statistics of the game, one group of variables directly associated to the match, non-sport variables, and finally those related to the referee. Among all the considered variables in the different groups, the following were chosen for the econometric models. In the game statistics category, the shots on target, for both home and away teams were labeled as "HST" and "AST", respectively. It seems to be reasonable that the number of shots on goal are involved in the result of a match. The number of fouls for both teams were "HF" and "AF". Finally, the yellow and red cards labelled as "HYR" and "AYR" for the home and away team, were also considered. One match variable was introduced and was introduced and defined as "DERBY", which represents a match played between teams from the same city or region, or between the strongest teams of the competition. This variable takes the value 1 if the match respond to a derby and zero, otherwise.

Variables considered as non-sport were those concerning the team's budgets, defined as the logarithm value of the home team budget, "BUDH", and "BUDA" the logarithm of the away team budget. Finally, variables related to the referee were the international referee experience, "INTERNATIONAL", which was scored as 1 if the referee had such experience, and 0 otherwise, and the logarithm of the number of years of experience in the Spanish first division, namely as "AGEXP". The logarithm of the referee's age is denoted by "AGEREF". A brief description of these variables is shown in Table 2.

In order to check the goodness of the fitting, we have calculated the following information criterium and statistics. The Akaike information criterium (AIC), the mean absolute error (MAE) and the root mean square error (RMSE) statistics are obtaining by

$$\text{AIC} = 2k - 2\ell_{\max},$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y}_i|,$$

$$\text{RMSE} = \left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2\right)^{1/2},$$

$$\text{Raw residuals} = y_i - \widehat{y}_i,$$

where k is the number of parameters of the model and $\ell_{\max}$ is the maximum value of the log-likelihood function.

Table 2: *Description of the variables.*

Variable	Description
Game statistics	
HST	Home team shots on target
AST	Away team shots on target
HF	Number of home team fouls.
AF	Number of away team fouls.
HYR	Home team yellow and red cards.
AYR	Away team yellow and red cards.
Match variable	
DERBY	Match played between teams from the same city or region or between the strongest teams in the league.
Extra games	
BUDH	Logarithm of home team budget
BUDA	Logarithm of away team budget
Referee	
INTERNATIONAL	Scored as 0 if the referee has no international experience and 1 if he does.
AGEXP	Logarithm of years of experience in the first division
AGEREF	Logarithm of referee's age

The results both under the standard linear regression model and under the proposed regression model are shown in Table 3. As we expected, the home and away shots on target are significant factors with the expected signs and at the 1% level of significance. Furthermore, the OLS model only detects another significant factor, namely, the home team's budget at the 5% significance level. The AIC is equal to 1194.69 and the MAE and RMSE statistics are 1.008 and 1.182, respectively. The estimations of the proposed model, in addition to finding the same results as the previous one, detect an important new factor concerning the referee subject: the fact that a referee is international reduces the expected points of the home team at the 10% significance level. In this sense, we can see that the "home effect" is lower in those matches in which there is an international referee. The AIC for the proposed model is 691.218 and the MAE and RMSE statistics are 0.921 and 1.115, respectively, i.e, these values are notably lower than the ones obtained for the standard linear model.

Figure 8 shows the raw residuals of the OLS and the proposed models (left plot) and box-and-whisker chart of the raw residuals (right plot). Both plots remark a greater dispersion of the OLS residuals.

Table 3: *Estimation results for the OLS and the proposed regression models.*

	OLS			Proposed model		
Variables	$\hat{\beta}$	Standard Error	p-value	$\hat{\beta}$	Standard Error	p-value
Intercept	2.286^{**}	1.156	0.049	2.153	1.822	0.238
DERBY	0.139	0.164	0.397	0.203	0.292	0.487
HST	0.130^{***}	0.024	0.000	0.249^{***}	0.047	0.000
AST	-0.175^{***}	0.028	0.000	-0.307^{***}	0.052	0.000
HF	0.004	0.016	0.816	-0.012	0.028	0.657
AF	0.014	0.015	0.371	0.015	0.027	0.565
HYR	-0.0652	0.045	0.151	-0.115	0.077	0.139
AYR	-0.039	0.042	0.345	-0.059	0.072	0.409
BUDH	0.129^{**}	0.072	0.075	0.235^{*}	0.138	0.089
BUDA	-0.111	0.069	0.109	-0.166	0.117	0.156
INTERNATIONAL	-0.220	0.159	0.165	-0.519^{*}	0.281	0.065
ACIENT	0.068	0.177	0.702	0.059	0.050	0.238
AGEREF	-0.017	0.031	0.593	-0.047	0.043	0.276
AIC		1194.69			691.218	
MAE		1.008			0.921	
RMSE		1.182			1.115	

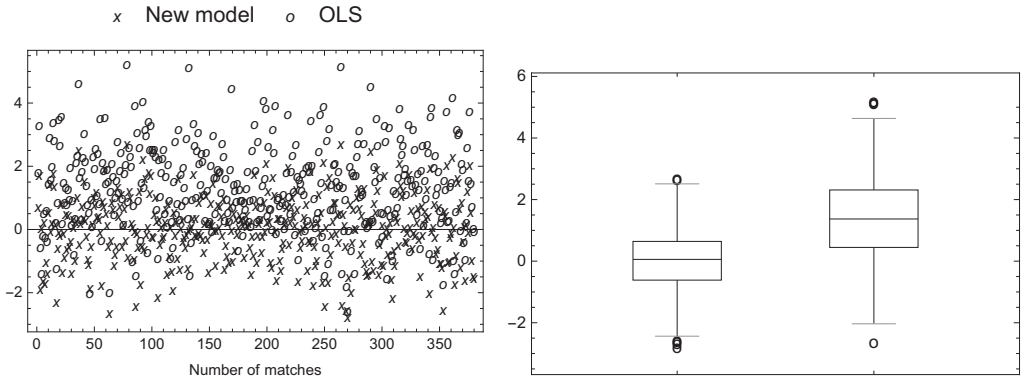

Figure 8: *Raw residuals and the box-and-whisker chart of the OLS (right) and proposed (left) raw residuals.*

Figure 9 illustrates the normal probability plot of the OLS residuals. As we can observe, the residuals plot is approximately linear supporting the condition that the error terms are normally distributed.

4. Discussion of results

In this study, we propose a new pmf for modeling the number of points achieved by a home or visitor team in a football match. In this context, we have analysed the number of points, firstly, without including covariates and, secondly, including this information.

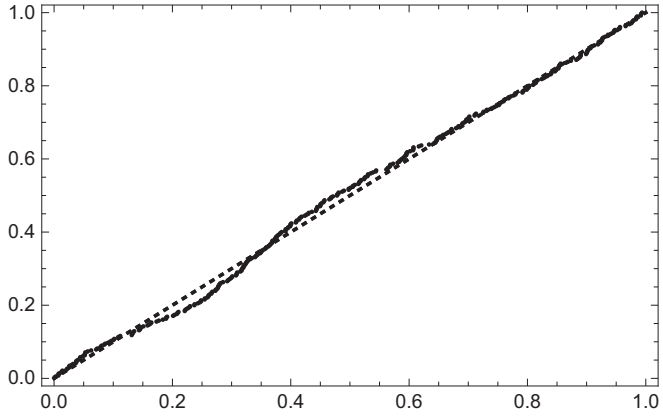

Figure 9: *Probability plot of the OLS residuals.*

The first part of the study includes estimation and prediction of the points achieved by the teams without using covariates. We can observe that this model provides good results except for some of the worst teams, i.e., the model fails only for two teams, Elche and Osasuna. While Elche finally remained in the First Division, Osasuna relegated. The second part focuses on the introduction of covariates. Several variables are considered from the Spanish Football League (First division) during the 2013-14 season and two sets of models are analysed. First, the home linear (standard) regression model. And second, the home regression model proposed in this paper. The results obtained indicate that the latter model produces better fits than the other standard model. In view of the good results obtained in this case and in the estimation without covariates previously studied, we believe the distribution given in (3) is appropriate for this data set.

Several factors, proposed in earlier studies, were considered relevant to the expected number of points. Karlis and Ntzoufras (2000) considered the number of goals as an indicator for the strength of a team and it can be used to determine the performance of a team. Rue and Salvesen (2006) ignored data as number of near goals, corners or free kicks and focused on the defending and attacking skills of each team. With the proposed model, home and away team shots on target, which are in some way the attacking strength, are the most significant factors considering the game statistics variables. With respect to the coefficients of the budgets, the large budgets home teams have more expected points. Finally, a significant factor appears regarding the referee variable which has not been significant in the previous models, namely, the international issue. International referees have a negative (positive) relationship with the home (away) teams indicating that this kind of referee is not influenced by the "home effect" of the match. Quite the opposite, if there is an international referee in the match, the expected number of points increases (decreases) for the away (home) team. These findings support those obtained in Pérez-Sánchez, Gómez-Déniz and Dávila-Cárdenes (2018) in which the authors proposed a skewed logistic model for estimating the probability of an away victory.

To conclude, in this work, significant variables are obtained not only related to the game itself, but also to the referees or even the economic potential of the teams. This fact may be used by others actors around the sport of football as coaches or even bookmakers, who analyse all the available information as part of their bets.

Appendix

Second partial derivatives to get the Fisher's information matrix are:

$$\frac{\partial^2 \ell(\tilde{x};\boldsymbol{\beta})}{\partial \beta_j^2} = \sum_{i=1}^{n} \frac{\partial^2}{\partial \beta_j^2} \kappa(\theta(\mu_{ij})) + \sum_{i=1}^{n} \frac{x_{ij}}{\theta(\mu_{ij})} \left[\frac{\partial^2}{\partial \beta_j^2} \theta(\mu_{ij}) - \left(\frac{1}{\theta(\mu_{ij})} \frac{\partial}{\partial \beta_j} \theta(\mu_{ij}) \right)^2 \right],$$

$$\frac{\partial^2 \ell(\tilde{x};\boldsymbol{\beta})}{\partial \beta_j \partial \beta_l} = \sum_{i=1}^{n} \frac{\partial^2}{\partial \beta_j \beta_l} \kappa(\theta(\mu_{ij})) + \sum_{i=1}^{n} \frac{x_{ij}}{\theta(\mu_{ij})} \left[\frac{\partial^2}{\partial \beta_j \beta_l} \theta(\mu_{ij}) - \left(\frac{1}{\theta(\mu_{ij})} \frac{\partial}{\partial \beta_j} \theta(\mu_{ij}) \right)^2 \right],$$

for $j = 1,2,\ldots,q$, $l = 1,2,\ldots,q$ and $j \neq l$ and the derivatives needed are the followings,

$$\frac{d\kappa(\varphi(\mu))}{d\varphi(\mu)} = -\frac{36 + 18\varphi(\mu)^2}{24 + \varphi(\mu)(6 + \varphi(\mu)^2)},$$

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{1}{3} x_j \mu_i \exp\left(\beta^\top \boldsymbol{x}\right),$$

$$\frac{d\varphi(\mu_i)}{d\mu_i} = \varphi(\mu_i) \left[\frac{\sqrt[3]{2}\left(4 - 2\mu_i - 2/3\psi(\mu_i)^{-1/3}\psi'(\mu_i)\right)}{\sqrt[3]{2}\left(3 - \mu_i(\mu_i - 4) - \psi(\mu_i)^{2/3}\right)} - \frac{\psi'(\mu_i)}{3\psi(\mu)} - \frac{1}{\mu_i - 3} \right],$$

$$\frac{d\psi(\mu_i)}{d\mu_i} = \mu_i \left[2(\mu_i - 3) + 4\mu_i + \frac{\sqrt{2(\mu_i - 3)}(\mu_i - 1)(2 + 19\mu_i(\mu_i - 3))}{\sqrt{\mu_i\left(3 + 19\mu_i(\mu_i - 3) - 1\right)}} \right].$$

Acknowledgement

EGD and JMPS work was partially funded by grant ECO2013–47092 (Ministerio de Economía y Competitividad, Spain and ECO2017–85577–P (Ministerio de Economía, Industria y Competitividad. Agencia Estatal de Investigación)).

References

Baio, G. and Blangiado, M. (2010). Bayesian hierarchical model for the prediction of football result. *Journal of Applied Statistics*, 37, 253–264.

Brillinger, D. (2008). Modelling game outcome of the Brazilian 2006 series a championship as ordinal-valued. *Brazilian Journal of Probability and Statistics*, 22, 89–104.

Díaz, I. and Núñez, V. (2010). On the use of simulation methods to compute probabilities: application to the first division Spanish soccer league. *SORT*, 34, 181–200.

Fisher, R. (1934). The effects of methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics*, 6, 13–25.

Greenhough, J., Birch, P., Chapman, S. and Rowlands, G. (2002). Football goal distributions and extremal statistics. *Physica A*, 316, 615–624.

Harandi, S.S. and Alamtsaz, M. (2013). Discrete alpha-skew-Laplace distribution. *SORT*, 39, 71–84.

Hon, L. and Parinduri, R. (2016). Does the three-point rule make soccer more exciting? evidence from a regression discontinuity design. *Journal of Sports Economics*, 17, 377–395.

Johnson, N., Kemp, A. and Kotz, S. (2005). *Univariate Discrete Distributions* John Wiley, INC.

Karlis, D. and Ntzoufras, I. (2000). On modelling soccer data. *Student*, 3, 229–244.

Karlis, D. and Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society. Series D, The Statistician*, 52, 133–145.

Khatri, C. (1959). On certain properties of power-series distributions. *Biometrika*, 46, 486–490.

Lehmann, E. and G. Casella, G. (1998). *Theory of Point Estimation* Springer, New York.

Louzada, F., Suzuki, A. and Salasar, L.B. (2014). Predicting match outcomes in the English premier league: Which will be the final rank? *Journal of Data Science*, 12, 235–254.

Noack, A. (1950). A class of random variables with discrete distributions. *The Annals of Mathematical Statistics*, 21, 127–132.

Patil, G. and Rao, C. (1978). Weighted distributions and size biased sampling with applications to wildlife populations and human families. *Biometrics*, 34, 179–184.

Pérez-Sánchez, J.M., Gómez-Déniz, E. and Dávila-Cárdenes, N. (2018). A comparative study of logistic models using an asymmetric link: Modelling the away victories in football. *Symmetry*, 10, 1–12.

Rue, H. and Salvesen, O. (2006). Prediction and retrospective analysis of soccer matches in a league. *Journal of the Royal Statistical Society. Series D. The Statistician*, 49, 399–418.

Ruskeepaa, H. (2009). *Mathematica Navigator. Mathematics, Statistics, and Graphics. Third Edition* Academic Press. USA.

Saraivaa, E., Suzuki, A., Ciro, A. and Luzadab, F. (2016). Predicting football scores via Poisson regression model: applications to the National Football League. *Communications for Statistical Applications and Methods*, 23, 297–319.

Wolfram, S. (2003). *The Mathematica Book* Wolfram Media, Inc.