Discrete Weibull regression for modeling football outcomes

Alessandro Barbiero

Department of Economics, Management and Quantitative Methods, Università di Milano, via Conservatorio 7, 20122 Milan, Italy, e-mail: alessandro.barbiero@unimi.it Phone: +390250321533, Fax: +3902503111

Abstract

We propose the use of the discrete Weibull distribution for modeling football match results, as an alternative to existing Poisson and generalized Poisson models. The number of goals scored by the two teams playing a football match are regarded as a pairwise observation and are modelled first through two independent discrete Weibull variables, and then through two dependent discrete Weibull variables, using a copula approach that accommodates non-null correlation. The parameters of the bivariate discrete Weibull distributions are assumed to depend on covariates such as the attack and defense abilities of the two teams and the 'home effect'. Several discrete Weibull regression models are proposed and then applied to the 2015-2016 Italian Serie A. Even if the interpretation of parameters is less immediate than in the case of bivariate Poisson models, nevertheless these models represent a suitable alternative, which can be applied also in other fields than sport data analysis.

Keywords: count data, count regression model, Frank copula, Poisson distribution, sport analytics

1. Introduction

Sports is a thriving field for applying statistical models and methods when dealing with athletic data, especially football outcomes. In the last decades, increasing sum of money has been put into circulation by sports related industries. On the one hand, sports clubs have become well organized companies, often present on the stock exchange, investing large amounts of money every year for equipping competitive teams. On the other hand, betting on the outcome of football matches (either the final score or other specific characteristics or events occurring during the match, such as half-time result, the scorer of the first goal, etc.) have attracted a lot of sport fans, even in countries lacking a long betting tradition, testified by a steady growth of on-line betting market (Calvosa, 2015). In making bets, the challenge is to seek for matches whose probabilities have been specified inaccurately, so that the expected gain is high.

Over the years, sports-betting has attracted the interest of scientists and, specifically, economists and statisticians. On the one side, scientists can help managers to make crucial decisions in certain circumstances and provide guidance that is not available without scientific search in large datasets; on the other side, they can help bookmakers calibrate the probabilities of game events and betters take advantage from poorly determined probabilities.

Lots of works have been published about the analysis and forecasting of sport results, developing several different statistical models and methods, covering a multitude of sports,

e.g. American football, baseball, ice hockey, and soccer, which is by far the most followed sport in many European countries.

When modeling football results, namely the final number of goals scored by the two competing teams in a football match, the Poisson distribution and related Poisson regression models have received great popularity, especially for their ease of use and interpretation. Since the unique parameter λ of the Poisson distribution corresponds to its expected value, λ assumes the natural meaning of 'scoring ability' of a team and can be regarded as a function of other observed variables. The simple bivariate independent Poisson model has been first used in football data analysis, and later more complex models allowing for a non-null correlation have been explored, since real data often show a slight but non-negligible positive correlation between the goals scored by the two teams. Important results about such methodologies can be found for example in Maher (1982); Lee (1997); Dixon and Coles (1997); Dyte and Clark (2000); Baio and Blangiardo (2010). In Karlis and Ntzoufras (2009), as an alternative to Poisson models for predicting the exact outcome of a game, the Skellam distribution is used to model the difference in the goals scored by the two teams. Most of the models developed in these works are 'static', in the sense that they do not take into account evolution of team abilities over time, i.e., throughout the considered season or championship; this simplifying assumption allows the use of random variables only to model the data. More recently, Koopman and Lit (2015) developed a dynamic bivariate Poisson model for analysing and forecasting match results, whose parameters are let change stochastically over time; in this case, stochastic processes are involved.

Despite their simplicity, basic Poisson-based models present some theoretical and practical weaknesses, namely 1) impossibility of dealing with over-dispersed data, i.e. data whose variance is larger than the mean, and 2) difficulty in dealing with excess of 0-0 draws, whose probability is usually under-estimated. These two drawbacks have been faced by using generalized Poisson models or diagonally-inflated Poisson models, which consider an inflation factor for diagonal terms (e.g., the $(0, 0)$ pair) in the bivariate joint distribution (Karlis and Ntzoufras, 2003, 2011). These models, although usually presenting a better flexibility and adaptability with respect to the original ones, lack a prompt and incisive interpretation of all their parameters.

In this work, we select and apply the discrete Weibull model, introduced by Nakagawa and Osaki (1975), for modeling the number of goals scored by the two competing teams in each match of a football round-robin tournament. The discrete Weibull random variable, which can be seen as a generalization of the geometric distribution, is characterized by two parameters, the first of which has an easy interpretation, and allows to model under-, equi-, and over-dispersed count data. Here we consider the discrete Weibull regression model described in Kalktawi et al (2015), derived as a discrete analogue of the more popular continuous Weibull regression model, often used in survival analysis. The discrete Weibull random variables modeling the number of goals scored by the two teams in a match will be modelled first through a simple independent bivariate model, and then through a model allowing for dependence, by resorting to a copula approach. Very recently, a bivariate Weibull count model has been indeed proposed in Boshnakov et al. (2017), based on 1) the probability distribution of the number of events occurring in a count process driven by independent and identically distributed Weibull inter-arrival times (also known as a Weibull renewal process), and 2) a copula linking two univariate count Weibull distributions together. The univariate

probability mass function of the Weibull count model is there expressed as an infinite sum, derived as a Taylor expansion of the continuous Weibull probability density function; it must not be confused with the discrete Weibull distribution in Nakagawa and Osaki (1975).

Through these models, we aim at finding results about team performances that the simple (final) rank table may mask. The objective is twofold: from the betters and sport fans' side, the statistical model represents a basis for calibrating probabilities for future outcomes; from the coaching and management staff's side, it can work as a decision support tool in order to highlight strengths and weaknesses of the team (and competitors) in order to deploy corrective drives.

The remaining of the paper is organised as follows. Section 2 introduces the discrete Weibull distribution as a discrete counterpart of the continuous Weibull distribution. Section 3 describes a general discrete Weibull regression model, which is declined in several variants, also including the use of copulas for controlling dependence. In Section 4, these models are applied to model soccer data and in Section 5 they are practically implemented to the Italian football championship data. Concluding remarks can be found in Section 6.

2. The type I discrete Weibull distribution

The discrete Weibull distribution was introduced by Nakagawa and Osaki (1975) as a discrete counterpart of the continuous Weibull distribution and is usually referred to as 'type I discrete Weibull distribution', in order to distinguish it from two other models proposed later by Stein and Dattero (1984) (type II discrete Weibull) and Padgett and Spurrier (1985) (type III discrete Weibull).

It is well known that a continuous Weibull r.v. T has probability density function (p.d.f.) given by

$$f_t(t;\lambda,\beta) = \lambda\beta t^{\beta-1}\mathrm{e}^{-\lambda t^{\beta}} \quad t>0,$$

with $\lambda,\beta>0$, and cumulative distribution function (c.d.f.)

$$F_t(t;\lambda,\beta) = 1-\mathrm{e}^{-\lambda t^{\beta}}. \tag{1}$$

If we consider the r.v. $Y=\lfloor T \rfloor$, where $\lfloor T \rfloor$ denotes the largest integer equal to or smaller than T, it can be easily shown that its probability mass function (p.m.f.), defined on the non-negative integers only, is given by

$$p(y;q,\beta) = F_t(y+1)-F_t(y) = \mathrm{e}^{-\lambda y^{\beta}}-\mathrm{e}^{-\lambda(y+1)^{\beta}} = q^{y^{\beta}}-q^{(y+1)^{\beta}} \quad y\in\mathbb{N}_0, \tag{2}$$

with $q=\mathrm{e}^{-\lambda}$, where $0<q<1$. For this discrete distribution (henceforth, simply 'discrete Weibull'), which was proposed by Nakagawa and Osaki (1975), the c.d.f. is

$$F(y;q,\beta) = 1-q^{(y+1)^{\beta}} \quad y\in\mathbb{N}_0. \tag{3}$$

We write $Y\sim \mathrm{DW}(q,\beta)$ to indicate that the r.v. Y follows the discrete Weibull distribution in (3) with parameters q and β. This distribution, differently from the two competitors cited before, retains the expression of the cumulative distribution function of the continuous Weibull model –just compare Eq.(1) to Eq.(3). Note that the first parameter q has a nice

interpretation: since $P(X = 0) = 1-q$, it represents the probability of a positive value. As to the second parameter β, it does not possess an equally nice interpretation. However, defining $S(y) = P(Y > y) = 1 - P(Y \leq y) = 1 - F(y)$ as the survival function and $r(y) = p(y)/S(y)$ as the hazard rate function of Y, it has been shown (Nakagawa and Osaki, 1975) that $r(y)$ is a constant function if $\beta = 1$ (equal to $(1-q)/q$; in this case, the DW distribution reduces to the geometric distribution), an increasing function if $\beta > 1$, a decreasing function if $\beta < 1$. Figure 1 displays the p.m.f. of the discrete Weibull r.v. for several value combinations of q and β. Here the role of β, for a fixed value of q, is clearer: larger values of β lead to less dispersed distributions, with most of the probability mass concentrated on the first integer values; smaller values of β lead to more dispersed distributions, with a thick right tail. The expected value of the discrete Weibull r.v. cannot be generally computed in a closed form; it is equal to the infinite sum:

$$\mathbb{E}(Y) = \sum_{y=1}^{\infty} q^{y^\beta}, \tag{4}$$

which leads to a closed expression if and only if $\beta = 1$: $\mathbb{E}(Y) = q/(1-q)$. From Eq.(4), it is clear that the expected value of a discrete Weibull r.v., fixed q, is a decreasing function of β. The expression in Eq.(4) can be approximated resorting to the expected value $\mathbb{E}(T)$ of the corresponding continuous distribution; Khan Khalique and Abouammoh (1989) showed that

$$\mathbb{E}(Y) < \mathbb{E}(T) = \left(-\frac{1}{\log q}\right)^{\frac{1}{\beta}} \Gamma\left(1 + \frac{1}{\beta}\right) < \mathbb{E}(Y) + 1$$

which ensures the value $\mathbb{E}(Y)$ falls between $\mathbb{E}(T) - 1$ and $\mathbb{E}(T)$. For example, for $q = 0.9$ and $\beta = 1.5$, we have that $\mathbb{E}(T) = 4.047$, and then $3.047 < \mathbb{E}(Y) < 4.047$; actually, by approximating the infinite series in Eq.(4), we derive $\mathbb{E}(Y) \approx 3.550$.

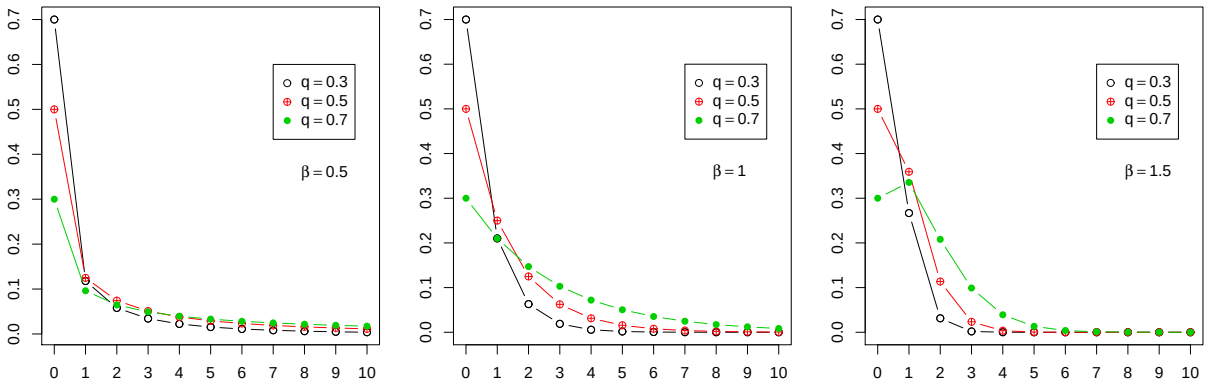


Figure 1: Graphs of probability mass function of the DW distribution for several value combinations of its parameters q and β

Contrary to the Poisson r.v., which cannot adequately model count data whose variance differs from the mean, a circumstance often occurring in practice, the discrete Weibull r.v. can model both under-dispersed and over-dispersed data (Englehardt and Li, 2011). This distribution can also handle count data presenting an excess of zeros, arising in many physical

situations (see again Englehardt and Li (2011)); just remember that the probability of 0 is controlled by the q parameter only.

As for point and interval estimation of the parameters of the discrete Weibull distribution, one can refer to Khan Khalique and Abouammoh (1989); Kulasekera (1994); Barbiero (2016), where several inferential procedures are considered: the standard maximum likelihood and moments' method (and modifications thereof) and the method of proportion; applicability issues and statistical properties are discussed.

The discrete Weibull model is implemented in the `R` environment (R Development Core Team, 2016) through the packages `DiscreteWeibull` (Barbiero, 2015) and `DWreg` (Vinciotti, 2015).

3. Discrete Weibull regression

In count regression models, where the dependent variable Y takes non-negative integer values, the dependence of the conditional mean $\mathbb{E}[Y_i|\boldsymbol{x}_i] = \mu_i$ on the covariates $\boldsymbol{x}_i$ is usually specified via the equation

$$g(\mu_i) = \boldsymbol{\alpha}'\boldsymbol{x}_i \tag{5}$$

where $g(\cdot)$ is a known link function and $\boldsymbol{\alpha}$ is the vector of regression coefficients, which are typically estimated by maximum likelihood. For the discrete Weibull model, specifying the dependence of the conditional mean of the dependent r.v. Y as in Eq.(5) is not viable, since its expected value does not have a closed analytical expression, see Eq.(4). A simple regression model for count data based on the discrete Weibull distribution was introduced by Kalktawi et al (2015) and implemented in the package `DWreg` (Vinciotti, 2015). First, one can see that the parameter q of the discrete Weibull distribution is equivalent to $\mathrm{e}^{-\lambda}$ in the continuous Weibull case, see Eq.(1). Since continuous Weibull regression imposes a link between the parameter λ and the covariates, at first the discrete Weibull regression was analogously introduced via the parameter q. For $i = 1, \ldots, n$, the response variable Y_i is assumed to have a conditional distribution given by $p(y_i, q(\boldsymbol{x}_i), \beta|\boldsymbol{x}_i)$, where $q(\boldsymbol{x}_i)$ is the first parameter of the discrete Weibull model related to the explanatory variables $\boldsymbol{x}_i$ through a complementary log-log link function: $\log(-\log(q_i)) = \boldsymbol{\alpha}'\boldsymbol{x}_i$. The complementary log-log function of q has the interval $(0, 1)$ as domain and takes values in $(-\infty, +\infty)$; its inverse is given by $q_i = \mathrm{e}^{-\mathrm{e}^{\boldsymbol{\alpha}'\boldsymbol{x}_i}}$. Thus, the p.m.f. of the dependent count variable Y_i can be rewritten as

$$p(y_i|\boldsymbol{x}_i) = \left(e^{-e^{\boldsymbol{\alpha}'\boldsymbol{x}_i}}\right)^{y_i^{\beta}} - \left(e^{-e^{\boldsymbol{\alpha}'\boldsymbol{x}_i}}\right)^{(y_i+1)^{\beta}},$$

where β is supposed to be constant. As an alternative to complementary log-log function, one can choose different transformations to link the dependent count variable Y with the set of covariates, for example, logit or probit.

Additionally, even the second parameter β can be related to explanatory variables $\boldsymbol{z}_i$, not necessarily the same as for q, through the following natural link function (remember that β takes only positive values): $\log(\beta_i) = \boldsymbol{\gamma}'\boldsymbol{z}_i$, which reversed returns $\beta_i = e^{\boldsymbol{\gamma}'\boldsymbol{z}_i}$. In this case, the p.m.f. of the dependent count variable Y_i takes the form

$$p(y_i|\boldsymbol{x}_i, \boldsymbol{z}_i) = \left(e^{-e^{\boldsymbol{\alpha}'\boldsymbol{x}_i}}\right)^{y_i^{e^{\boldsymbol{\gamma}'\boldsymbol{z}_i}}} - \left(e^{-e^{\boldsymbol{\alpha}'\boldsymbol{x}_i}}\right)^{(y_i+1)^{e^{\boldsymbol{\gamma}'\boldsymbol{z}_i}}}.$$

For this general model, given a sample $y_1, \ldots, y_n$, in order to obtain the MLEs for the unknown vectors of parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, the log-likelihood function, which can be written as

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\gamma}; y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) = \log \prod_{i=1}^{n} p(y_i | \boldsymbol{x}_i, \boldsymbol{z}_i) = \sum_{i=1}^{n} \log p(y_i | \boldsymbol{x}_i, \boldsymbol{z}_i), \qquad (6)$$

can be numerically maximized using standard optimization tools; for example, in the `R` environment, the function `maxLik` of the homonym package.

4. Regression models for modeling football outcomes

In this section, we delineate some discrete Weibull regression models that can be used for modeling football match results, with particular reference to round-robin tournaments.

4.1. Italian Serie A

We will focus on the main Italian football championship, called Serie A, a professional league competition for football clubs located at the top of the Italian football league system. For most of Serie A's history, there were 16 or 18 clubs competing; since 2004-05, there have been $N = 20$ clubs altogether. As in most of the European countries (e.g. England, Germany, Spain,...), in Italy a true round-robin format is used. During the league, from August to May, each club plays each of the other teams twice; once at home and once away, totaling 38 games by the end of the season. In the first half of the season, called the 'andata', each team plays once against each league opponent, for a total of 19 games. In the second half of the season, called the 'ritorno', the teams play in exactly the same order that they did in the first half of the season, the only difference being that home and away situations are switched. Since the 1994-95 season, teams are awarded three points for a win, one point for a draw and no points for a loss.

For the Italian Serie A, an overall number of $n = 380$ games are played during the season. Here we are interested in analysing and modeling the final result of each game. For the game i, $1 \leq i \leq 380$, we denote with y_{1i} the number of goals scored by the home team, h_i, and with y_{2i} the number of goals scored by the away team, a_i. We assume that y_{1i} and y_{2i} can be marginally modelled as discrete Weibull r.v.s. According to whether y_{1i} and y_{2i} can be assumed to be independent and according to which parameters are modelled through a regression model (and with which set of covariates) several bivariate count models arise. We are presenting them in the following subsections.

4.2. Independence models

For the models of this subsection, we assume that y_{1i} and y_{2i} can be modelled as independent discrete Weibull r.v.s. Differences from one model to another stand the choice of the parameters to be regressed and the sets of covariates.

4.2.1. First model

In the first model, we employ a regression model for the first parameter q only of the two discrete Weibull distributions. We assume also that the first parameter of the distribution of scored goals depends on the attack strength of the team playing and on the defense strength

of the opposing team, and that there is also a 'home effect', that is an advantage for the team playing 'at home'. Formally, we assume $Y_{1i} \sim \mathrm{DW}(q_{1i}, \beta)$, $Y_{2i} \sim \mathrm{DW}(q_{2i}, \beta)$ with

$$\begin{aligned} \log[-\log(q_{1i})] &= \mu^{(q)} + \mathrm{home}^{(q)} + \mathrm{att}_{h_i}^{(q)} + \mathrm{def}_{a_i}^{(q)} \\ \log[-\log(q_{2i})] &= \mu^{(q)} + \mathrm{att}_{a_i}^{(q)} + \mathrm{def}_{h_i}^{(q)} \end{aligned} \tag{7}$$

for $i = 1, 2, \ldots, n$, where n is the number of games bivariate observations, i is a game or observation indicator, h_i and a_i indicate the home and away team in game i, respectively; $\mu^{(q)}$ is a constant parameter (the intercept); $\mathrm{home}^{(q)}$ is the home effect parameter, $\mathrm{att}^{(q)}$ and $\mathrm{def}^{(q)}$ encapsulate the offensive and defensive performances of the team. As suggested for example in Karlis and Ntzoufras (2003), to achieve identifiability of the above model parameters, we use sum-to-zero constraints for ease of interpretation, so that the constant parameters $\mu^{(q)}$ specify q_1 and q_2 when two teams of the same strength play on a neutral field:

$$\sum_{j=1}^{N} \mathrm{att}_j^{(q)} = \sum_{j=1}^{N} \mathrm{def}_j^{(q)} = 0$$

Offensive and defensive parameters, due to the sum-to-zero constraint, are expressed as departures from a team of average offensive or defensive ability.

4.2.2. Second model

Compared to the first one, this model further assumes that the shape parameters β of the discrete Weibull distributions depend upon a set of covariates. We assume that this set is exactly the same as for the first parameter q: also β, for the distribution of the number of goals scored by a given team, depends on the attack strength of the team, on the defense strength of the opposing team, and on the 'home effect'. We assume $Y_{1i} \sim \mathrm{DW}(q_{1i}, \beta_{1i})$, $Y_{2i} \sim \mathrm{DW}(q_{2i}, \beta_{2i})$ with

$$\begin{aligned} \log[-\log(q_{1i})] &= \mu^{(q)} + \mathrm{home}^{(q)} + \mathrm{att}_{h_i}^{(q)} + \mathrm{def}_{a_i}^{(q)} \\ \log[-\log(q_{2i})] &= \mu^{(q)} + \mathrm{att}_{a_i}^{(q)} + \mathrm{def}_{h_i}^{(q)} \end{aligned} \tag{8}$$

$$\begin{aligned} \log[\beta_{1i}] &= \mu^{(\beta)} + \mathrm{home}^{(\beta)} + \mathrm{att}_{h_i}^{(\beta)} + \mathrm{def}_{a_i}^{(\beta)} \\ \log[\beta_{2i}] &= \mu^{(\beta)} + \mathrm{att}_{a_i}^{(\beta)} + \mathrm{def}_{h_i}^{(\beta)} \end{aligned} \tag{9}$$

Again, as suggested in Karlis and Ntzoufras (2003), to achieve identifiability of the above model parameters, we use sum-to-zero constraints for ease of interpretation, so that the two constant parameters $\mu^{(q)}$ and $\mu^{(\beta)}$ specify q_1 and q_2, and β_1 and β_2, respectively, when two teams of the same strength play on a neutral field.

$$\sum_{j=1}^{N} \mathrm{att}_j^{(q)} = \sum_{j=1}^{N} \mathrm{def}_j^{(q)} = \sum_{j=1}^{N} \mathrm{att}_j^{(\beta)} = \sum_{j=1}^{N} \mathrm{def}_j^{(\beta)} = 0$$

Offensive and defensive parameters, due to the sum-to-zero constraint, are expressed as

departures from a team of average offensive or defensive ability.

4.2.3. Third model

The third model considers the same regression model for the q parameter as for the first and second models, and a different set of covariates for the regression of the shape parameter β of the distribution of scored goals, which is supposed to depend only on the scoring team.

$$\begin{aligned} \log[-\log(q_{1i})] &= \mu^{(q)} + \text{home}^{(q)} + \text{att}_{h_i}^{(q)} + \text{def}_{a_i}^{(q)} \\ \log[-\log(q_{2i})] &= \mu^{(q)} + \text{att}_{a_i}^{(q)} + \text{def}_{h_i}^{(q)} \end{aligned} \tag{10}$$

$$\begin{aligned} \log(\beta_{1i}) &= \mu^{(\beta)} + \text{team}_{h_i}^{(\beta)} \\ \log(\beta_{2i}) &= \mu^{(\beta)} + \text{team}_{a_i}^{(\beta)} \end{aligned} \tag{11}$$

This model is more parsimonious and seems to be 'more reasonable' than the second model, since the shape parameter β plays a 'minor' and less direct role in determining the expected value of the discrete Weibull distribution, whereas it more considerably affects its dispersion, which can be assumed to depend on the team only, and not on its specific attack/defense abilities, on the team it is playing against, or on the field where the game is played. Again, we impose regression parameters have to satisfy a sum-to-zero constraint:

$$\sum_{j=1}^{N} \text{att}_j^{(q)} = \sum_{j=1}^{N} \text{def}_j^{(q)} = \sum_{j=1}^{N} \text{team}_j^{(\beta)} = 0.$$

4.3. Copula models

Lack of independence/incorrelation between the number of goals scored by the two teams in a football match was first claimed by Dixon and Coles (1997); in McHale et al (2011) the use of copulas for modeling the two correlated distributions was first suggested. Here we assume that the random variables modeling the number of goals scored by home and away teams, Y_{1i} and Y_{2i}, are no longer statistically independent, given the covariates; we model their dependence structure through a specific copula family.

If we suppose Y_{1i} and Y_{2i} are discrete (Weibull) r.v.s., with c.d.f F_{1i} and F_{2i}, linked by a generic copula C, then their joint c.d.f. is given by $F(y_{1i}, y_{2i}) = C(F_{1i}(y_{1i}), F_{2i}(y_{2i}); \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is some vector (or scalar) parameter characterizing the copula, and the joint p.m.f. of the random pair (Y_{1i}, Y_{2i}) can be computed by using the relationship between bivariate p.m.f. and c.d.f.:

$$P(Y_{1i} = y_{1i}, Y_{2i} = y_{2i}) = F(y_{1i}, y_{2i}) - F(y_{1i} - 1, y_{2i}) - F(y_{1i}, y_{2i} - 1) + F(y_{1i} - 1, y_{2i} - 1), \tag{12}$$

where Y_{1i} and Y_{2i} are marginally distributed according to the first, second or third regression model discussed in Section 4. This way, other three possible models arise, corresponding to the first, second and third independence model (which we can denote as the fourth, fifth and sixth model, respectively).

From among the multitude of parametric bivariate copulas, we pick Frank's copula, be-

longing to the so-called Archimedean family. One-parameter Frank copula is defined as

$$C(u_1, u_2) = -\frac{1}{\kappa} \ln \left[1 + \frac{(e^{-\kappa u_1} - 1)(e^{-\kappa u_2} - 1)}{e^{-\kappa} - 1} \right], \quad (u_1, u_2) \in (0, 1)^2,$$

with $\kappa \neq 0$. Frank copula, through the choice of its parameter κ, allows both negative ($\kappa < 0$) and positive ($\kappa > 0$) dependence. As a limiting case, for $\kappa \to 0$, we have $\lim_{\kappa \to 0} C(u_1, u_2; \kappa) = u_1 u_2$, and thus Frank copula reduces to the independence copula; for $\kappa \to \infty$, Frank copula tends to the upper Fréchet bound or comonotonicity copula; for $\kappa \to -\infty$, Frank copula tends to the lower Fréchet bound or countermonotonicity copula. This means that Frank copula interpolates between perfect positive and negative dependence. Moreover, it is able to capture weaker tail-dependence better than Gaussian copula; it is expected to work satisfactorily if linking the number of goals scored by the two competing teams, which are usually characterized by a slight (often positive, sometimes negative) correlation. It has been already satisfactorily employed by Boshnakov et al. (2017).

The values of the κ parameter can be better interpreted resorting to the expression of Spearman's correlation for Frank copula (valid however for continuous margins only; see Nelsen (1999)):

$$\rho^S = 1 - \frac{12}{\kappa} \left(\frac{1}{\kappa} \int_0^\kappa \frac{a}{e^a - 1} \mathrm{d}a - \frac{2}{\kappa^2} \int_0^\kappa \frac{a^2}{e^a - 1} \mathrm{d}a \right) \tag{13}$$

The plot of $\rho^S(\kappa)$ is reported in Figure 2.

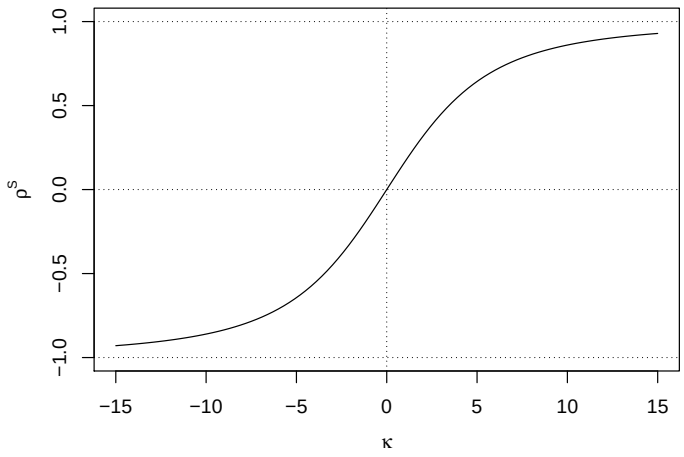


Figure 2: Spearman's ρ for Frank copula as a function of the parameter κ.

For this dependence model, the marginal parameters related to q_1, β_1, q_2, β_2, and the copula parameter κ, which is assumed (for the sake of simplicity) to be constant, can be simultaneously estimated by resorting to the customary maximum likelihood method, with an additional computational cost.

5. Application to Italian Serie A

In this section, we apply the models introduced and described in the previous section to the Italian Serie A data for the 2015-2016 season.

Some tests were preliminary performed to check whether the number of goals scored by the two opponent teams are actually independent. The standard test for assessing whether Pearson's correlation ρ is equal to zero (against the two-sided alternative hypothesis) has a sample value -0.062 with a p-value of 23%. Spearman's sample rank correlation ρ^S is equal to -0.033 and the test used for assessing if Spearman's rho is equal to zero against the two-sided hypothesis (rank correlation independence test) has a p-value equal to 52%. The two tests confirm previous findings that there is no evidence of dependence between the variables modeling the number of goals scored by the two opponent teams in each game. So, even if in some models, through the use of copulas, we will take into account the dependence between the two variables, as a first step it is reasonable to adopt the bivariate independent models of Section 4.2.

Based on the sample data (y_{1i}, y_{2i}), $i = 1, \ldots, n = 380$, for the first three models we compute the maximum likelihood estimates of the regression parameters, by using the `R` package `DWreg`; for the last three models we properly modified the code in `DWreg` in order to obtain the maximum likelihood estimates of the regression parameters and of the copula parameter κ simultaneously. Once the regression parameters are computed, for each game i, we compute the values of q_{1i}, q_{2i}, β (or β_{1i} and β_{2i}) and then derive the marginal distributions of the Y_{1i} and Y_{2i} r.v.s. and then compute - resorting to the independence condition between Y_{1i} and Y_{2i} (for the first three models) or to Eq.(12) (for the last three models) - the probability of any possible outcome for each match and, properly aggregating these probabilities, the probability of a win, a draw or a loss for the home or away teams. By considering all the n games one can then reconstruct also the overall number of scored and conceded goals and the overall number of points predicted by the model for each team at the end of the championship, and then build a predicted final scoreboard, which can be compared with the actual final scoreboard in order to qualitatively establish to what extent the model is satisfactory. A quantitative assessment of the reliability of each model is performed by computing usual goodness-of-fit indexes as log-likelihood, AIC, and BIC. All the computations and simulations have been carried out in the `R` programming environment; the relevant code is available on request.

5.1. First model

Estimated attack and defense parameters for the first model are reported in Table 1. Note that - recalling the complementary log-log function used as link function for the q parameter - a negative value (and higher in absolute value) of the attack parameter estimate leads to a larger value of the estimate of q_{1i} and then indicates a better attack (or, better, a larger probability of scoring); vice versa, a positive value (and higher in absolute value) of the defense parameter estimate leads to a smaller value of the estimate of q_{2i} and indicates a better defense (or, better, a smaller probability of conceding goals). Note that the (attack/defense) parameter estimates for the last team (in alphabetical order, Verona) have been computed so that the sum of the 20 team estimates is zero (zero-sum constraint).

To better understand the meaning of attack and defense parameter estimates, in Table 1, in the last two columns, a complementary log-log transformation is reported: $\exp(-\exp(\mu^{(q)}+$

$\text{att}^{(q)}$)) and $\exp(-\exp(\mu^{(q)} + \text{def}^{(q)})$. The transformed 'attack' ('defense') parameters can be interpreted as probabilities of scoring (conceding) at least one goal playing away against a team with average defense (attack) ability. Having the best attack parameter does not necessarily mean having the best attack (i.e., being the team that scored most goals), but simply having the largest probability of scoring (recall the considerations about q in Section 2). Analogous interpretation holds for the defense parameter. From the table, it can be noticed that Roma is the team with the best attack parameter, Palermo that with the worst one; Juventus is the team with the best defense parameter, Frosinone that with the worst one. These results returned by the model are quite in-line with actual data. Roma scored at least one goal in 34 out of 38 games (the best result, together with Juventus); Bologna in 20 out of 38 games (the worst result; Palermo scored at least one goal in 23 games). Juventus conceded at least one goal in 16 games out of 38 (the best performer); Verona in 35 games (Frosinone in 31).

Table 1: Parameter estimates (columns 2 and 3) and their transformations (columns 4 and 5) for the 'first' bivariate discrete Weibull model for 2015-2016 Italian Serie A data. Significance codes for p-values: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

team	attack	defense	transf.attack	transf.defense
Juventus	-0.558***	0.961***	0.816	0.395
Napoli	-0.709***	0.402*	0.840	0.588
Roma	-0.720***	0.142	0.841	0.664
Inter	0.079	0.224	0.681	0.641
Fiorentina	-0.329*	0.214	0.775	0.644
Sassuolo	-0.021	0.217	0.706	0.643
Milan	0.004	0.100	0.700	0.676
Lazio	-0.138	-0.092	0.734	0.723
Chievo	0.128	0.066	0.668	0.685
Empoli	0.186	0.028	0.652	0.694
Genoa	0.123	-0.052	0.669	0.714
Atalanta	0.223	0.086	0.642	0.679
Torino	-0.126	-0.132	0.731	0.733
Bologna	0.322.	0.102	0.613	0.675
Sampdoria	-0.032	-0.286.	0.709	0.766
Palermo	0.339*	-0.379*	0.608	0.784
Udinese	0.285.	-0.390*	0.624	0.786
Carpi	0.324.	-0.241	0.612	0.757
Frosinone	0.285.	-0.612***	0.624	0.825
Verona	0.335	-0.357	0.609	0.780
Other parameters		*Goodness-of-fit*		
$\mu^{(q)}$	-1.036***	$\ell_{\max}$	-1032.325	
$\text{home}^{(q)}$	-0.383***	AIC	2146.65	
$\hat{\beta}$	1.864***	BIC	2336.616	

Moving to the total number of goals, Roma is the team who scored more goals, Palermo

scored 5 goals more than Bologna, who has the actual worst attack. Juventus is the team who conceded least goals (20); Frosinone had actually the worst defense (76 goals conceded).

From the estimates in Table 1, it can be also deduced that an average team playing away against a team of the same strength would have a q attack parameter equal to $\exp(-\exp(-1.036)) = 0.701$, that means it would score at least one goal the 70% of cases; an average team playing at home against a team of the same strength would have a q attack parameter equal to $\exp(-\exp(-1.036 - 0.383)) = 0.785$, that would mean it would score at least one goal the 78.5% of cases. Note that, as expected, there is a 'positive' home effect, i.e., playing at home increases the probability of scoring at least one goal.

Figure 3: Attack vs. defense effects for each of the 20 teams of Italian Serie A 2015-2016, calculated according to the first regression model. The two series of parameters can be interpreted as follows: on the x-axis, the probability of scoring at least one goal playing away with a team with average defense ability; on the y-axis, the probability of conceding at least one goal playing away with a team with average attack ability.

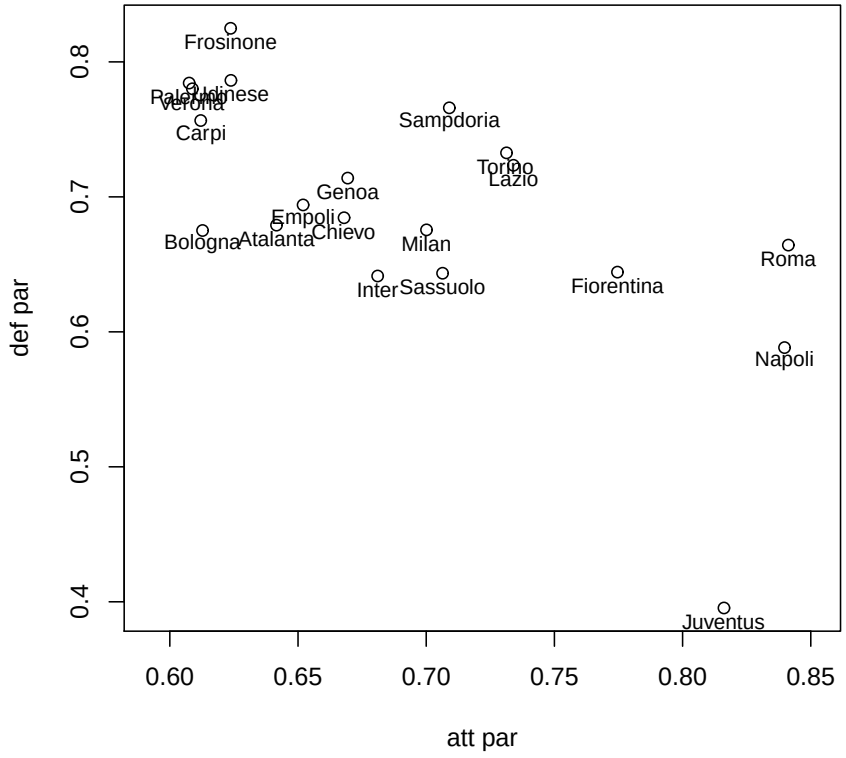


Figure 3 shows the (average) transformed attack vs. defense effects for each of the 20 teams. On the right side of the bi-plot, the teams with larger probability of scoring; on the bottom side, the teams with smaller probability of conceding goals. Note that the team that won the 2015-2016 serie A season, Juventus, is by far the team with the best defense, and with an attack slightly less powerful than Napoli and Roma. Note also that the teams that acquired

the last five positions in the final rank (Verona, Frosinone, Carpi, Palermo and Udinese) occupy the top-left corner in the biplot (Palermo and Verona are almost undistinguishable) and are actually quite distant from the other teams. The use of the 'transformed' parameters instead of original regression parameters is preferable: since they are probabilities, they are easy to understand and can be directly compared. The β parameter estimate, computed equal to 1.864, does not influence the probability of scoring, controlled by the q parameter, but jointly with q, models the entire distribution of scored goals and then its average value. Note that the estimate value, largely greater than 1, indicates - as expected - that the goal distribution is practically concentrated on the first integers (i.e., the probability of a team scoring, say, 10 goals, is almost null). For the 'average team' (playing away or at home against a team of the same strength), the probability distribution and related expected values are reported in Table 2.

Table 2: Distribution of the number goals scored by an 'average' team playing against a team of the same strength according to the first regression model

	0	1	2	3	4	≥ 5	expected value
away	0.299	0.426	0.211	0.055	0.008	0.001	1.050
home	0.215	0.371	0.261	0.113	0.033	0.007	1.402

Table 1 reports also the maximum value of the log-likelihood function $\ell_{\max} = -1032.325$ for the model at study, as well as two other indices, which can be used for comparing different models; the Akaike Information Criterion, AIC $= 2k - 2\ell_{\max}$, and the Bayesian Information Criterion, BIC $= k \log n - 2\ell_{\max}$: the smaller the values of these two indices, the better the fit of the model to data.

For comparison purposes, we also fitted the bivariate independent Poisson model to these data, and the corresponding value of AIC is 2148.6, whereas for the bivariate independent discrete Weibull model is 2146.65. Thus the latter fits better than the former.

Table 3 displays the expected number of scored and conceded goals for each of the 20 teams, as well as the final expected number of points, computed according to the estimated model, compared with the actual values. The final scoreboard was calculated as follows: for each match i, the goal distributions of the home and away teams, Y_{1i} and Y_{2i}, were computed based on the discrete Weibull regression model. Then, the probability of winning, drawing and losing were computed for the home team, given by $p_{wi} = P(Y_{1i} > Y_{2i})$, $p_{di} = P(Y_{1i} = Y_{2i})$ and $p_{li} = P(Y_{1i} < Y_{2i})$, respectively. The expected number of points gained by the home team for match i is then given by $3p_{wi} + p_{di}$; for the away team, $p_{di} + 3p_{li}$. The expected number of goals scored or conceded by a team for each match is easily computed resorting to Eq.(4). Replicating this computation for each of the 380 matches leads to the final expected scoreboard. Comparing the actual and expected number of scored (or conceded) goals, we note that differences are 'reasonably' small and thus the model seems to fit the data satisfactorily. However, moving to the number of points gained by each team, which is indeed a much more important value to predict, since it determines the final ranking, we note that there are few apparent discrepancies between actual and theoretical values of these quantities. Just to cite the three most notable cases, on the one hand, Inter's number of expected points (55.45) is much smaller than the number of points actually gained (67): a

Table 3: Actual and expected number of goals and points for the 'first' bivariate discrete Weibull model. GF=goals for; GA=goals against

	actual			expected		
	GF	GA	points	GF	GA	points
Juventus	75	20	91	72.75	20.31	86.54
Napoli	80	32	82	79.45	33.57	79.55
Roma	83	41	80	79.38	41.37	74.33
Inter	50	38	67	45.28	40.34	55.45
Fiorentina	60	42	64	60.96	39.94	65.99
Sassuolo	49	40	61	48.80	40.40	57.92
Milan	49	43	57	47.69	44.29	54.46
Lazio	52	52	54	52.58	50.83	53.62
Chievo	43	45	50	43.36	45.67	50.38
Empoli	40	49	46	41.39	47.06	47.99
Genoa	45	48	46	43.33	49.85	47.72
Atalanta	41	47	45	40.30	45.10	48.41
Torino	52	55	45	52.05	52.37	52.33
Bologna	33	45	42	37.28	44.73	46.23
Sampdoria	48	61	40	48.22	58.73	46.12
Palermo	38	65	39	35.97	63.48	34.88
Udinese	35	60	39	37.57	63.83	35.91
Carpi	37	57	38	36.66	57.56	38.30
Frosinone	35	76	31	37.10	74.31	31.11
Verona	34	63	28	36.13	62.50	35.46

gross explanation maybe the singular behaviour of Inter in the first part of the season, which saw the team win several games with the minimum effort, i.e., by 1-0. On the other hand, Torino and Verona number of points are heavily overestimated by the model (52.33 vs 45 for Torino; 35.46 vs 28 for Verona); in particular, Verona ended the season at the bottom of the table, but the statistical model ranks them third to last.

The prediction of the number of draws is a problem often reported by researchers for Poisson models (Maher, 1982; Lee, 1997), which has been overcome by diagonal-inflated models (Karlis and Ntzoufras, 2003). Fitted counts for draws are shown in Table 4. Note that the model considerably underestimates the number of 0-0 draws and considerably overestimates the number of 1-1 draws. The actual overall number of draws is 95 versus an expected number of 97.3.

Table 4: Expected and actual frequencies of draws for the first model

result	0-0	1-1	2-2	3-3	$\geq$4-4	tot.
exp.freq.	24.85	53.73	16.64	1.95	0.12	97.3
act.freq.	31	44	15	5	0	95

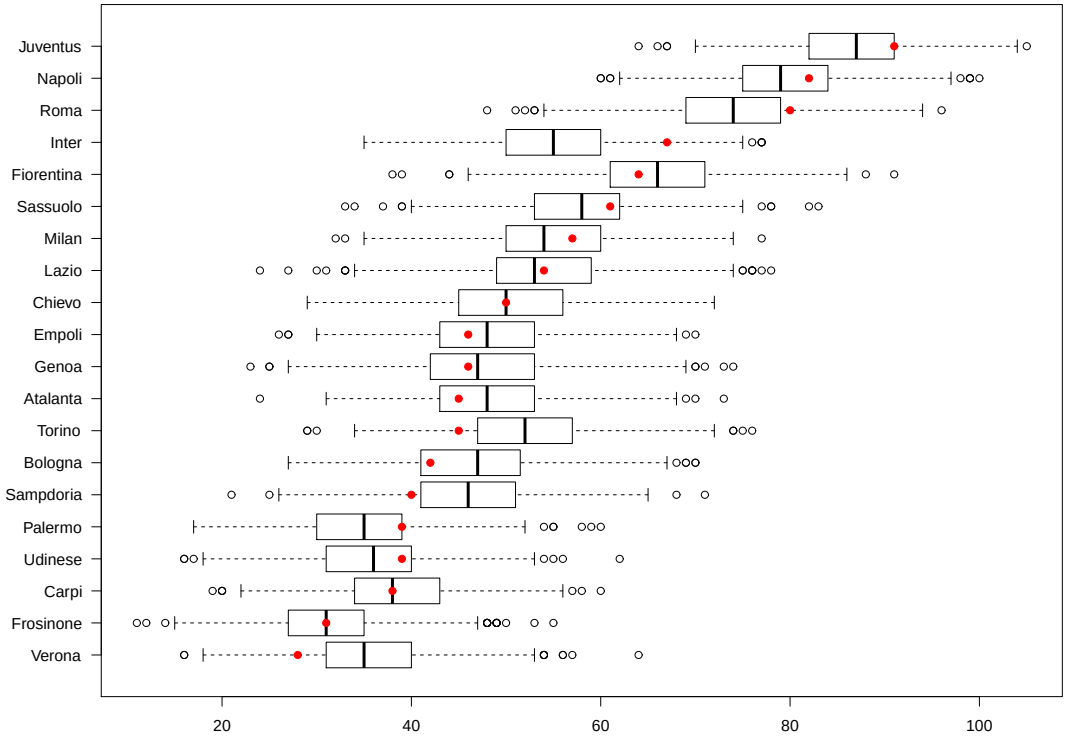
In Figure 4, the results are reported of a Monte Carlo simulation, consisting of $1,000$ simulated championships based on the parameter estimates of Table 1. For each team, the boxplot of the Monte Carlo distribution of the number of points gained over the $1,000$ artificial championships is displayed. Superimposed red points indicate the actual number of points of each team at the end of 2015-2016 season. Note that the teams whose actual number of points lies outside the corresponding box are Inter, Torino, Verona (as expected) plus Sampdoria and Roma.

In the next two sub-sections, we assess if the proposed bivariate discrete Weibull model may benefit by relaxing the condition of constance for β, imposing the regression model of Eq.(9) and thus introducing additional parameters.

5.2. Second model

We consider the case where β is regressed over the same set of covariates considered for q – Eq.(9) – so that now we have to estimate another set of attack and defense parameters. Table 5 reports the parameter estimates and their transformations for the 'second' bivariate discrete Weibull model. It is interesting to compare the values of the transformed attack and defense parameters related to q moving from Table 1 and Table 5, and how the values of the transformed attack and defense parameters related to β for the second model (last and second-to-last columns) differ from the common unique estimate of β of the first model. Differences between the attack and defense parameters related to q look, at a first glance, non-negligible; having introduced regressors also for β has lead to a stronger diversification among teams. However, as we told in the introduction and in the previous sub-section, interpreting the different values of β is a much more difficult task than interpreting different values of q. Comparison between β attack (or defense) parameters could be made between different teams for a same (or very similar) value of the q attack (defense) parameter. For example, Lazio and Udinese have very close values for the transformed attack parameter (0.639 and 0.645), but the values of the transformed attack parameter for β are quite different. Since

Figure 4: Boxplots of MC distribution of the number of points achieved by the 20 teams of Italian Serie A 2015-2016 computed over $1,000$ simulations according to the first model (independence model). The red circles indicate the actual number of points.



this value is much smaller for Lazio, it indicates a more dispersed distribution of the number of scored goals, and then a larger expected value. Roma and Sampdoria have a very similar value for the q transformed defense parameter (0.777 and 0.786), but the β transformed defense parameters are very different (2.500 and 1.870), indicating that the distribution of the number of conceded goals is much more concentrated on the first integers for Roma, leading to a smaller expected value.

Focusing on Roma team, we can show some calculations for illustrative purposes. For the first model the transformed attack parameter related to q was 0.841 and the transformed defense parameter to 0.664. The estimate of β was 1.864, common to all teams. Now, in the second model, the transformed attack parameter increases to 0.930 and the transformed defense parameter 0.777 (denoting a larger probability of scoring and a larger probability of conceding at least one goal playing away against an average team). But also the transformed attack and defense parameters related to β have increased, moving to 2.555 and 2.500, respectively: this increase somehow balance the increase in the q parameter, avoiding the mass of probability scattering on the right tail (remember the role of β hinted at Section 2). Note that now the estimate of the shape parameter β for a team of average strength playing away against a team of the same strength is $\exp(0.582) = 1.800$; if it plays at home, the estimates is 1.960 (in the first model, the estimate of β was 1.864.)

The number of scored and conceded goals and points predicted by the model for each team at the end of the championship are reported in Table 6. This second model seems to overcome most of the miss-matches of the first one. The largest difference in absolute value between expected and actual number of points is now 7.1 (for Inter).

The maximum value of the log-likelihood function is obviously larger than for the first model. However, looking at the values of AIC and BIC, it emerges that this model has actually a worse fit than the first one: the addition of new regression parameters does not increase the value of the log-likelihood function sensibly.

5.3. Third model

We consider now the regression model of Eq.(11), where the parameter β for the r.v. modeling the number of goals scored by a team is assumed to depend only on the team itself. In simpler terms, now each team j is characterized by a shape parameter β_j. Using the maximum likelihood method as usual, we derive the parameter estimates, which are reported in Table 7. The corresponding expected number of goals scored and conceded and the expected number of points for each team are reported in Table 8.

As for the previous model, let us focus on Roma team. Now, the transformed attack and defense parameters of the regression model for its q parameter are 0.919 and 0.674, meaning that when playing away with a team of average defense and attack strength, the probability of scoring at least one goal is 0.919 and the probability of conceding at least one goal is 0.674. The estimate of the β parameter is 2.438, which is, by the way, the largest value among all the 20 teams. The smallest value of β (1.443) belongs to Bologna, which has also the smallest value for q and the worst actual attack. From the data, one can see that in 18 games out of 38, it did not score, but it was able to score 2 goals in 7 games and 3 goals in 3; which demonstrates its scoring ability, though scarce, is characterized by some 'picks'.

As expected, the maximum value of the log-likelihood function lies between those corresponding to the first and second models. Looking at the AIC and BIC values, we notice that

Table 5: Parameter estimates and their transformation for the 'second' bivariate discrete Weibull model for 2015-2016 Italian Serie A data.
Significance codes for p-values: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

team	att.q	tr.att.q	def.q	tr.def.q	att.β	tr.att.β	def.β	tr.def.β
Juventus	-0.958*	0.874	0.987***	0.390	0.193	2.170	-0.034	1.730
Napoli	-0.705.	0.841	0.515*	0.556	0.022	1.829	-0.074	1.662
Roma	-1.571**	0.930	-0.330	0.777	0.356*	2.555	0.334*	2.500
Inter	-0.133	0.735	0.488*	0.564	0.143	2.065	-0.204	1.459
Fiorentina	-0.586	0.823	0.354	0.606	0.164	2.108	-0.106	1.609
Sassuolo	-0.328	0.777	-0.027	0.711	0.202	2.191	0.178	2.138
Milan	0.087	0.682	0.032	0.696	-0.009	1.774	0.064	1.908
Lazio	0.242	0.639	0.186	0.655	-0.211	1.449	-0.151	1.539
Chievo	0.429.	0.583	0.324	0.616	-0.183	1.491	-0.147	1.545
Empoli	-0.067	0.720	0.138	0.668	0.211	2.210	-0.062	1.683
Genoa	0.437	0.581	0.036	0.695	-0.181	1.493	-0.034	1.730
Atalanta	0.341	0.610	0.001	0.704	-0.083	1.647	0.060	1.901
Torino	-0.206	0.751	-0.490	0.807	0.082	1.943	0.193	2.169
Bologna	0.681**	0.500	0.305	0.621	-0.272	1.363	-0.135	1.563
Sampdoria	0.157	0.663	-0.376	0.786	-0.117	1.593	0.044	1.870
Palermo	0.559*	0.541	-0.177	0.745	-0.155	1.532	-0.109	1.605
Udinese	0.223	0.645	-0.192	0.748	0.094	1.965	-0.099	1.621
Carpi	0.421	0.586	-0.103	0.729	-0.039	1.721	-0.080	1.652
Frosinone	0.553*	0.543	-0.640.	0.831	-0.192	1.476	0.023	1.830
Verona	0.424	0.585	-1.031	0.882	-0.024	1.746	0.400	2.669

Other parameters	q	β	*Goodness-of-fit*	
μ	-1.047***	0.582***	$\ell_{\max}$	-1012.467
home	-0.513***	0.091	AIC	2184.934
			BIC	2500.148

Table 6: Actual and expected number of goals and points for the 'second' bivariate discrete Weibull model. GF=goals for; GA=goals against

	actual			expected		
	GF	GA	points	GF	GA	points
Juventus	75	20	91	74.12	20.96	89.92
Napoli	80	32	82	78.98	33.28	79.53
Roma	83	41	80	81.40	42.26	80.60
Inter	50	38	67	46.91	40.26	59.91
Fiorentina	60	42	64	60.83	40.16	68.15
Sassuolo	49	40	61	50.01	41.28	58.32
Milan	49	43	57	46.68	44.26	52.61
Lazio	52	52	54	51.39	49.25	52.70
Chievo	43	45	50	42.43	44.20	50.34
Empoli	40	49	46	41.98	46.75	50.29
Genoa	45	48	46	41.76	49.11	45.65
Atalanta	41	47	45	41.13	45.67	46.99
Torino	52	55	45	51.75	54.21	49.03
Bologna	33	45	42	36.74	44.67	45.16
Sampdoria	48	61	40	48.79	59.58	44.16
Palermo	38	65	39	35.77	63.36	35.89
Udinese	35	60	39	37.60	63.25	37.79
Carpi	37	57	38	36.45	57.82	38.84
Frosinone	35	76	31	36.55	73.64	29.97
Verona	34	63	28	35.38	62.68	29.36

this model has a fit still worse than the first one, but better than the second one.

Table 7: Parameter estimates and their transformation for the 'third' bivariate discrete Weibull model for 2015-2016 Italian Serie A data.
Significance codes for p-values: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

team	attack q	tr.attack q	defense q	tr.def. q	team β	tr. team β
Juventus	-0.999*	0.882	0.947***	0.417	0.200	2.271
Napoli	-0.657.	0.839	0.420*	0.597	-0.008	1.846
Roma	-1.386**	0.919	0.150	0.674	0.270*	2.438
Inter	-0.161	0.749	0.203	0.660	0.155	2.173
Fiorentina	-0.481	0.811	0.197	0.661	0.094	2.044
Sassuolo	-0.328	0.783	0.222	0.654	0.195	2.262
Milan	0.124	0.681	0.115	0.683	-0.049	1.771
Lazio	0.117	0.683	-0.092	0.734	-0.125	1.641
Chievo	0.391	0.605	0.073	0.694	-0.151	1.600
Empoli	0.004	0.711	0.028	0.705	0.150	2.162
Genoa	0.387	0.607	-0.037	0.721	-0.150	1.601
Atalanta	0.227	0.653	0.099	0.687	0.022	1.901
Torino	-0.141	0.745	-0.134	0.743	0.031	1.919
Bologna	0.658**	0.519	0.088	0.690	-0.254	1.443
Sampdoria	0.072	0.694	-0.289.	0.775	-0.042	1.784
Palermo	0.517*	0.566	-0.401*	0.797	-0.107	1.671
Udinese	0.271	0.641	-0.417*	0.799	0.036	1.928
Carpi	0.406	0.601	-0.269	0.771	-0.035	1.796
Frosinone	0.512*	0.568	-0.581***	0.827	-0.154	1.595
Verona	0.469	0.581	-0.322	0.782	-0.079	1.719

Other parameters	q	β	*Goodness-of-fit*	
μ	-1.080***	0.621***	$\ell_{\max}$	-1023.887
home	-0.361***	—	AIC	2167.774
			BIC	404.184

5.4. Copula models

In this section, we examine the results of the application of the copula model of Section 4.3 to the data. For the sake of brevity, we will discuss in depth only the fourth model. The parameter estimates are organized in Table 9. We will not dwell on the meaning of the marginal parameters' values, but we will rather focus on the copula parameter estimate, $\hat{\kappa} = 0.562$: Since it is a positive value, it denotes a positive dependence between the variables modeling the number of goals of the home and away teams. An approximate value of the corresponding Sperman's correlation, computed resorting to Eq.(13), which however assumes continuous margins, is 0.093. This result is quite unexpected, since we found out

Table 8: Actual and expected number of goals and points for the 'third' bivariate discrete Weibull model. GF=goals for; GA=goals against

	actual			expected		
	GF	GA	points	GF	GA	points
Juventus	75	20	91	74.93	20.70	90.96
Napoli	80	32	82	79.50	33.29	79.35
Roma	83	41	80	82.29	40.95	80.40
Inter	50	38	67	46.99	40.99	57.56
Fiorentina	60	42	64	61.67	40.53	67.15
Sassuolo	49	40	61	50.19	40.22	60.90
Milan	49	43	57	47.08	43.90	53.62
Lazio	52	52	54	51.10	50.95	51.08
Chievo	43	45	50	41.82	45.62	47.84
Empoli	40	49	46	42.10	47.10	49.30
Genoa	45	48	46	41.73	49.46	45.51
Atalanta	41	47	45	40.45	44.77	48.79
Torino	52	55	45	52.11	52.46	52.36
Bologna	33	45	42	36.09	45.41	42.92
Sampdoria	48	61	40	47.85	58.84	45.31
Palermo	38	65	39	34.92	64.38	33.36
Udinese	35	60	39	37.75	65.00	35.40
Carpi	37	57	38	36.31	58.70	37.23
Frosinone	35	76	31	36.22	72.58	30.85
Verona	34	63	28	35.68	60.94	35.54

that both Pearson's and Spearman's sample correlations between the number of goals scored by the home and away team were negative. However, those values were computed without considering any assumption on the underlying distributions and any regression model, i.e., are non-parametric estimates, so this discrepancy is somehow justified, also considering that the MLE of κ has an associated p-value 0.094: the hypothesis that the true κ is equal to zero is rejected at the 10% level, but is accepted at the 5% level. So actually (as anticipated by the independence tests performed) there is not a very significant dependence between the two variables Y_1 and Y_2, although the corresponding bivariate model is more adapt (in terms of AIC, but not in terms of BIC) than the analogous independence model to fit the data.

It is interesting to compare the values of the estimates of the other model parameters –those related to the scale parameter q and the shape parameter β of the marginal Weibull distributions– between Tables 1 and 9, i.e., before and after the hypothesis of independence was relaxed. The differences are, at a first sight, very small. Looking at their transformations, differences in absolute value are never greater than 0.0043, and are associated to the attack parameter of Carpi and the defense parameter of Bologna. The values of $\text{home}^{(q)}$ and $\mu^{(q)}$ differ for about 1 over 1000. The common value of $\hat{\beta}$ remains the same, at least at the third decimal digit.

It may be then interesting to assess, in practical terms, what is the effect of taking into account the dependence between the two count variables through the additional copula parameter. Let us consider the match Verona-Roma, whose final result was 1-1, and let us construct the probabilities of the possible outcomes by 1) considering the simple independence model (first model), 2) the copula model.

According to the bivariate independent model, by using the estimates in Table 1, we have that $Y_1 \sim \text{DW}(q_1 = \exp[-\exp(-1.036-0.383+0.335+0.142)] = 0.677, \beta_1 = 1.864)$ and $Y_2 \sim \text{DW}(q_2 = \exp[-\exp(-1.036 - 0.720 - 0.357)] = 0.885, \beta = 1.864)$. So we have, for example, that $P(Y_1 = 0) = 1 - 0.677 = 0.323$ and $P(Y_2 = 0) = 1 - 0.885 = 0.115$, and the probability of the outcome 0-0 is given by the joint probability $P(Y_1 = 0, Y_2 = 0) = 0.323 \cdot 0.115 = 0.037$.

According to the copula model, by using the estimates in Table 9, we have that $Y_1 \sim \text{DW}(q_1 = \exp[-\exp(-1.035-0.384+0.337+0.152)] = 0.674, \beta_1 = 1.864)$ and $Y_2 \sim \text{DW}(q_2 = \exp[-\exp(-1.035 - 0.715 - 0.368)] = 0.887, \beta = 1.864)$. So we have, for example, that $P(Y_1 = 0) = 1 - 0.674 = 0.326$ and $P(Y_2 = 0) = 1 - 0.887 = 0.113$, and the probability of the outcome 0-0 is given by the joint probability:

$$P(Y_1 = 0, Y_2 = 0) = -\frac{1}{0.562} \log \left\{ 1 + \frac{[\exp(-0.562 \cdot 0.326) - 1][\exp(-0.562 \cdot 0.113) - 1]}{\exp(-0.562) - 1} \right\} = 0.043$$

The two complete joint p.m.f.s are displayed in Table 10. The difference may be not clear at a first glance; however, moving from the independence model to the copula model, one can notice an increase of probability of the 0-0 and 0-1 outcomes, whereas 0-3, 0-4 and 0-5 outcomes, which were rare events and characterized by a large goal difference, further decrease their probability.

Does the copula model improve the prediction of draws? As we did in Section 5.1, we computed the number of 0-0, 1-1, etc. outcomes predicted by the model; results are reported in Table 11. The predicted number of 0-0 has now increased and is much closer to the real value (28.36 vs 31). The other predicted values slightly change; the predicted number of 1-1,

Table 9: Parameter estimates (columns 2 and 3) and their transformation (columns 4 and 5) for the fourth bivariate discrete Weibull model with Frank copula for 2015-2016 Italian Serie A data.
Significance codes for p-values: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 "" 1

team	att. q	def. q	transf.att.q	transf.def.q
Juventus	-0.564***	0.967***	0.817	0.393
Napoli	-0.701***	0.411*	0.838	0.585
Roma	-0.715***	0.152	0.840	0.661
Inter	0.068	0.229	0.684	0.640
Fiorentina	-0.330*	0.220	0.775	0.642
Sassuolo	-0.020	0.223	0.706	0.641
Milan	0.005	0.099	0.700	0.675
Lazio	-0.145	-0.098	0.735	0.725
Chievo	0.125	0.061	0.669	0.685
Empoli	0.184	0.030	0.652	0.693
Genoa	0.124	-0.063	0.669	0.716
Atalanta	0.227	0.088	0.640	0.678
Torino	-0.116	-0.129	0.729	0.732
Bologna	0.316.	0.085	0.614	0.679
Sampdoria	-0.035	-0.290.	0.710	0.766
Palermo	0.337*	-0.378*	0.608	0.784
Udinese	0.279	-0.392*	0.625	0.786
Carpi	0.337*	-0.227	0.608	0.753
Frosinone	0.287.	-0.620***	0.623	0.826
Verona	0.337	-0.368	0.608	0.782
Other parameters		*Goodness-of-fit*		
$\mu^{(q)}$	-1.035***	$\ell_{\max}$	-1030.94	
home$^{(q)}$	-0.384***	AIC	2145.88	
$\hat{\beta}$	1.864***	BIC	2340.479	
$\hat{\kappa}$	0.562.			

Table 10: Predicted outcomes for the match Verona-Roma; comparison between independence model with covariates for q (first model, a) and corresponding copula model (fourth model, b)

(a) independent bivariate model

VER-ROM	0	1	2	3	4	≥ 5	
0	0.037	0.078	0.081	0.061	0.037	0.028	0.323
1	0.050	0.105	0.110	0.083	0.049	0.038	0.435
2	0.022	0.047	0.049	0.037	0.022	0.017	0.193
≥ 3	0.006	0.012	0.012	0.009	0.006	0.004	0.049
	0.114	0.242	0.252	0.190	0.113	0.088	

(b) copula model

VER-ROM	0	1	2	3	4	≥ 5	
0	0.043	0.087	0.082	0.057	0.032	0.024	0.326
1	0.048	0.104	0.110	0.084	0.050	0.040	0.436
2	0.018	0.041	0.047	0.039	0.025	0.020	0.191
≥ 3	0.004	0.010	0.012	0.010	0.006	0.005	0.047
	0.113	0.241	0.252	0.190	0.114	0.089	

which was largely overestimated by the first model, decreases of about one half, but it is still far from the actual value. Overall, the predicted number of draws is larger than that provided by the first model and than the real one. As expected, it looks like the introduction of (positive) statistical dependence through Frank copula has overall 'inflated' the bivariate probabilities of diagonal elements.

Table 11: Expected and actual frequencies of draws for the copula model (fourth model)

result	0-0	1-1	2-2	3-3	$\geq$4-4	tot.
exp.freq.	28.36	53.20	17.62	2.29	0.61	102.09
act.freq.	31	44	15	5	0	95

We fit also the two copula models with extra regression parameters for β; Table 12 reports concise results. The AIC and BIC of the other two copula models are larger than the basic copula model. The additional complexity introduced through the extra regression parameters for β is not worth the increase in the log-likelihood function.

We can conclude that for the dataset at hand, a bivariate discrete Weibull model, with

Table 12: Summary results for copula-based models

model	n.par.	$\hat{\kappa}$	$\ell_{\max}$	AIC	BIC
4th (team att. def. for q)	41	0.562	-1030.94	2145.88	2340.479
5th (team att. def. for q and β)	81	0.663	-1010.641	2183.282	2558.581
6th (team att. def. for q and team for β)	61	0.575	-1022.464	2166.928	2449.56

correlated dependent components (the dependence being accommodated by Frank copula) and a set of covariates for the scale parameter q only, represents the best compromise in terms of parsimony and absolute goodness-of-fit.

6. Conclusions

Most of the models developed in the literature for analysing football outcomes employ the Poisson or some generalized Poisson univariate or bivariate distribution for modeling the number of goals scored by the two teams. The contribution of this work consists of proposing the Weibull distribution (which can be regarded as a generalized geometric distribution and has been so far employed in applications in the engineering and environmental sciences) as an alternative model that can be used also when the assumption of independence between the two variables does not hold true; in this case, a suitable copula (Frank copula, allowing for the entire range of dependence) can be introduced for linking the two variables together. The additional computational work for recovering the model parameters from the data is exiguous; their interpretation may be less immediate than the case of independent bivariate Poisson, but there may be a substantial benefit in terms of goodness of fit. The application in the Italian Serie A is provided as an illustrative and persuasive example. It shows how, from among several models, a bivariate dependent model with the scale parameter q of the Weibull variates depending on attack and defense ability of each team and home effect is the best choice in terms of AIC. Although the proposed model has been discussed from an exploratory perspective and not explicitly inflected to a predictive framework, we believe that it will be helpful for both systems players (in order to predict probability of outcomes more satisfactorily) and team managers (who can easily interpret regression parameters for their team and support them in undertaking subsequent policies for enhancing its performance).

Improvement of the proposed models can be pursued, for example looking at the many modified or generalized discrete Weibull distributions that have been recently made available in the literature. Another research perspective concerns the choice of the copula function linking the two count distribution. Here we adopted Frank copula, an Archimedean copula sharing most properties of the Gaussian copula (symmetry, radial symmetry, tail-independence, full range of dependence, etc). Other copula families can be explored, for example Farlie-Gumbel-Morgenstern, which has a very simple expression and allows only moderate level of correlations. Since the number of scored and conceded goals in football matches are typically characterized by a very small correlation, this family could represent a more viable alternative to capture dependence. Furthermore, focusing on the peculiar features of the data, overstepping measures of fit as customary AIC and BIC, which are based on the fit of all the bivariate sample observations (i.e., all the matches results of a championship), one can investigate for other global measures, which express the degree of matching between actual and predicted 'aggregated' results (total number of points earned by the teams, final ranking, etc.).

References

Baio, G., & Blangiardo, M. (2010). Bayesian hierarchical model for the prediction of football results, Journal of Applied Statistics, 37(2), 253–264.

Barbiero, A. (2015). DiscreteWeibull: Discrete Weibull Distributions (Type 1 and 3). R package version 1.1. `https://CRAN.R-project.org/package=DiscreteWeibull`

Barbiero, A. (2016). A comparison of methods for estimating parameters of the type I discrete Weibull distribution. Statistics and its Interface, 9(2), 203–212.

Boshnakov, G., Kharrat, T., & McHale, I. G. (2017). A bivariate Weibull count model for forecasting association football scores. International Journal of Forecasting, 33(2), 458-466.

Calvosa, P. (2015). On Consumer Behaviour and Customer On Line Purchasing Motivation: An Empirical Investigation into the Sports-Betting Industry in Italy. Economia dei Servizi, (1), 17–40.

Dixon, M. J., & Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. Journal of the Royal Statistical Society: Series C (Applied Statistics), 46(2), 265–280.

Dyte, D., & Clarke, S. R. (2000). A ratings based Poisson model for World Cup soccer simulation. Journal of the Operational Research Society, 51(8), 993–998.

Englehardt, J.D., & Li, R.C. (2011). The discrete Weibull distribution: An alternative for correlated counts with confirmation for microbial counts in water. Risk Analysis, 31, 370–381 (2011).

Kalktawi, H. S., Vinciotti, V., & Yu, K. (2015). A Simple and Adaptive Dispersion Regression Model for Count Data. arXiv preprint arXiv:1511.00634.

Karlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. Journal of the Royal Statistical Society: Series D (The Statistician), 52(3), 381–393.

Karlis, D., & Ntzoufras, I. (2009). Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference. IMA Journal of Management Mathematics, 20(2), 133–145.

Karlis, D., & Ntzoufras, I. (2011). Robust fitting of football prediction models. IMA Journal of Management Mathematics, 22, 171–182.

Khan, M.S.A., Khalique, A., & Abouammoh, A.M. (1989) On estimating parameters in a discrete Weibull distribution, IEEE Transactions on Reliability, 38(3), 348–350.

Koopman, S. J., & Lit, R. (2015). A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League. Journal of the Royal Statistical Society: Series A (Statistics in Society), 178(1), 167–186.

Kulasekera, K.B. (1984) Approximate MLE's of the parameters of a discrete Weibull distribution with type I censored data. Microelectronics Reliability, 34(7), 1185–1188.

Lee, A. J. (1997). Modeling scores in the Premier League: is Manchester United really the best?. Chance, 10(1), 15–19.

Maher, M. J. (1982). Modelling association football scores. Statistica Neerlandica, 36(3), 109–118.

McHale, I., & Scarf, P. (2011). Modelling the dependence of goals scored by opposing teams in international soccer matches. Statistical Modelling, 11(3), 219–236.

Nakagawa, T., & Osaki, S. (1975). The discrete Weibull distribution. IEEE Transactions on Reliability, 24(5), 300–301.

Nelsen, R.B. (1999). An Introduction to Copulas, Springer-Verlag: New York.

Padgett, W.J., & Spurrier, J.D. (1985). Discrete failure models. IEEE Transactions on Reliability, 34(3), 253–256 (1985).

R Development Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, `http://www.R-project.org/`.

Stein, W.E., & Dattero, R. (1984) A new discrete Weibull distribution. IEEE Transactions on Reliability, 33, 196–197.

Vinciotti, V. (2015). DWreg: Parametric Regression for Discrete Response. R package version 1.0. `https://CRAN.R-project.org/package=DWreg`