MODEL KONKURENČNÍCH RIZIK S NETRADIČNÍ APLIKACÍ

COMPETING RISKS MODEL WITH A NON-TRADITIONAL APPLICATION

Petr Volf

Adresa: UTIA AV ČR, Praha

E-mail: volf@utia.cas.cz

Abstrakt: Model konkurenčních rizik je aplikován na analýzu časů prvních gólů ve fotbalových zápasech. Konkurují si latentní časy vstřelení gólu obou týmů. Předpokládá se jejich exponenciální rozdělení, jejich vzájemná závislost je popsána pomocí vhodné kopuly. V příspěvku jsou zpracována data ze dvou ročníků české první fotbalové ligy, 2014–2016.

Klíčová slova: konkurenční rizika, analýza přežití, sportovní statistika.

Abstract: The competing risks scheme is used to the analysis of time to the first goal in a football (soccer) match. The competing random variables are two latent times (of two teams) to scoring. It is assumed that these times are exponentially distributed, their mutual dependence is described by a copula ensuring the model identifiability. As a real example the data from two seasons 2014–2016 of the Czech First League are analyzed and compared.

Keywords: competing risks, survival analysis, sports statistics.

1. Introduction and motivation

The main theme of the present study has arisen from the question how important is the first goal in the football match and, consequently, how strong is the effort of both teams to score first. The problem of the first goal impact has already been addressed in several papers, either directly, as in Nevo and Ritov (2013), or implicitly, in dynamic models of scoring process, as for instance in Dixon and Robinson (1998) and Volf (2009). The latter also contains a brief simulation study showing the dependence of final result on the first goal author and time, for teams with different characteristics. A common feature of such models is the dependence of scoring intensities not only on the characteristics of teams, but also on the match development. As expected, the results of such studies indicate that the reaction to the goal, either scored or obtained, can differ and depends on the teams strengths; not just on factual, but also on psychical strength.

A basic probabilistic model for the final score of a football match, proposed for instance in Maher (1982), consists of two conditionally independent Poisson random variables. More flexible models are obtained for instance by using inflated Poisson models, which are actually the mixtures of Poisson distribution with a discrete distribution giving higher weights to certain more frequent results (as 0:0, 1:0, 1:1). Another generalization can consist in considering changes or/and a time development of model parameters as well as covariates during the match (again for instance Dixon and Robinson, 1998, or Volf, 2009). Yet another way of the basic model improvement consists in considering an explicit form of dependence between both teams scoring distributions. Let us mention here at least two relevant papers. Karlis and Ntzoufras (2003) have employed a special case of the bivariate Poisson distribution. In the same context, McHale and Scarf (2011) have described the dependence with the aid of a copula model. Interesting is the comparison of conclusions of both approaches. While the correlation in the former model is non-negative (by definition), the latter paper concludes that the correlation is negative and is absolutely larger in more competitive matches. It has to be also said that the use of copula in discrete distribution models is not easy technically (and then computationally), because marginal distribution functions are as a rule expressed by sums of point probabilities, not having a reasonably closed form.

In the present contribution we concentrate to the analysis of the distribution of time to the first goal. Consequently, we deal with continuous-type distributions of random times, their dependence is described with the aid of the competing risks scheme. On one hand the use of a copula for two-dimensional continuous distribution can lead to a nicely closed form of the model, on the other hand it is well known that in the competing risks setting the model may not be identifiable. A proof and an example of this phenomenon are given in Tsiatis (1975), some instances of identifiable (or not) models are presented in Basu and Ghosh (1978) – in these classical studies the notion of copula has not been used yet. Therefore we are facing the problem of reasonable copula selection. Fortunately, it has been revealed (cf. Zheng and Klein, 1995) that when the main objective is to estimate the correlation of competing random variables, the selection of copula type is not of crucial importance.

The next section recalls briefly the scheme of competing risks and the problem of possible non-identifiability. The Barnett copula model is formulated and the way of its maximum likelihood evaluation is derived. The main body of the paper is devoted to the application to the analysis of real data from one season of the Czech First League. The results are discussed and compared with the data from another season, the impact of the first goal on the final match result is examined, too.

2. Competing risks model

The competing risks model assumes that certain events (e. g. a failure of a device) can be caused by K different reasons. Such a situation is then modelled by K (possibly dependent) random variables (random times, as a rule) $T_j,\ j = 1, \ldots, K$, sometimes accompanied by a variable C of random right censoring. C is then independent of all T_j. Let $\overline{F}_K(t_1, \ldots, t_K) = P(T_1 > t_1, \ldots, T_K > t_K)$ be the joint survival function of $\{T_j\}$. However, instead of the net times T_j we standardly observe just crude data (sometimes called also the identified minimum) $Z = \min(T_1, \ldots, T_K, C)$ and the indicator $\delta = j$ if $Z = T_j,\ \delta = 0$ if $Z = C$.

2.1. Problem of identifiability

The data structure described above allows a direct estimation of the distribution of $Z = \min(T_1, \ldots, T_K)$, for instance its survival function $S(t) = P(Z > t) = \overline{F}_K(t, \ldots, t)$. Further, we can estimate so called incidence densities

$$f_j^*(t) = \mathrm{d}\, P(Z = t, \delta = j) = -\frac{\partial \overline{F}_K(t_1, \ldots, t_K)}{\partial t_j}\Big|_{t_1 = \ldots = t_K = t}$$

and also their integrals, the cumulative incidence functions, often called also the crude distribution functions. They are estimable consistently by standard survival analysis methods, see for instance Lin (1997). However, in general, from data (Z_i, δ_i), $i = 1, \ldots, N$, it is not possible to identify either marginal or joint distribution of $\{T_j\}$. Tsiatis (1975) has shown that for an arbitrary joint model we can find a model with independent components having the same incidences, i.e. we cannot distinguish between the models. Namely, this independent model is given by cause-specific hazard functions $h_j^*(t) = f_j^*(t)/S(t)$. As a consequence of the Tsiatis (1975) result, it is necessary to make certain functional form assumptions about the type of both marginal and joint distribution in order to identify them. Several such cases are studied in Basu and Ghosh (1978) and in some other papers. More recent results on identifiability can be found for example in Schwarz et al. (2011) dealing with non-parametric setting, or in Escarela and Carriere (2003) considering Frank copula and parametric models.

2.2. Copula models for dependence

In the sequel we shall consider just 2 competing events, i.e. random variables S, T and data $Z_i = \min(S_i, T_i, C_i)$, $\delta_i = 1, 2, 0$, $i = 1, \ldots, N$. The notion of copula offers a way how to model multivariate distributions, we prefer here to use it for modelling the joint survival function $\overline{F_2}(s,t)$ of S, T:

$$\overline{F_2}(s,t) = P(S > s, T > t) = C(\overline{F_S}(s), \overline{F_T}(t), \theta), \tag{1}$$

$\overline{F_S}$, $\overline{F_T}$ are marginal survival functions of S and T, $C(u, v, \theta)$ is a copula, i.e. a two-dimensional distribution function on $[0,1]^2$, with uniformly distributed marginals U, V. θ is a copula parameter, which is, as a rule, uniquely connected with the correlation of U and V, hence also with the correlation of S and T. It is seen that the use of a copula allows us to model the dependence structure separately from the analysis of marginal distributions. From this point of view, the identifiability of the copula (and its parameter) and of marginal distributions can be considered as two separate steps. Zheng and Klein (1995) proved that when the copula is known, the marginal distributions are estimable consistently (and then the joint distribution, too, from (1)), even in a non-parametric (so quite general) setting. However, in general, the value of θ has to be known. The problem of proper copula choice is analyzed in a set of papers. Let us mention here Kaishev et al. (2007) comparing performances of several copula types. As already mentioned, a common agreement is that the knowledge (or a good estimate) of parameter θ is much more crucial for a reasonable estimation of a joint distribution. As a consequence, because the knowledge of copula type is still an unrealistic supposition, we can try to use a sufficiently flexible class of copulas, as an approximation, and concentrate on reliable estimation of parameters.

2.3. Barnett copula

Tsiatis (1975) in his example considers two competing random variables S and T with exponential distributions and the following marginal and joint survival functions,

$$\overline{F}_S(s) = \mathrm{e}^{-\lambda s}, \;\; \overline{F}_T(t) = \mathrm{e}^{-\mu t}, \;\; \overline{F}_2(s,t) = \mathrm{e}^{-\lambda s - \mu t - \gamma st}\,.$$

The example actually uses so called Barnett copula

$$C(u,v) = u \cdot v \cdot \exp(-\theta \cdot \ln u \cdot \ln v), \tag{2}$$

where $\theta \geq 0$. It follows that when U, V are uniform on $[0,1]$ random variables tied by copula (2), their correlation $\rho(U,V) \leq 0$; $\theta = 0$ means independence of U and V. Further, when (2) is used for connection of S, T from above, the parameters are related in the following way: $\gamma = \theta \cdot \lambda \cdot \mu$. Figure 1 shows the dependence of correlation on parameters.

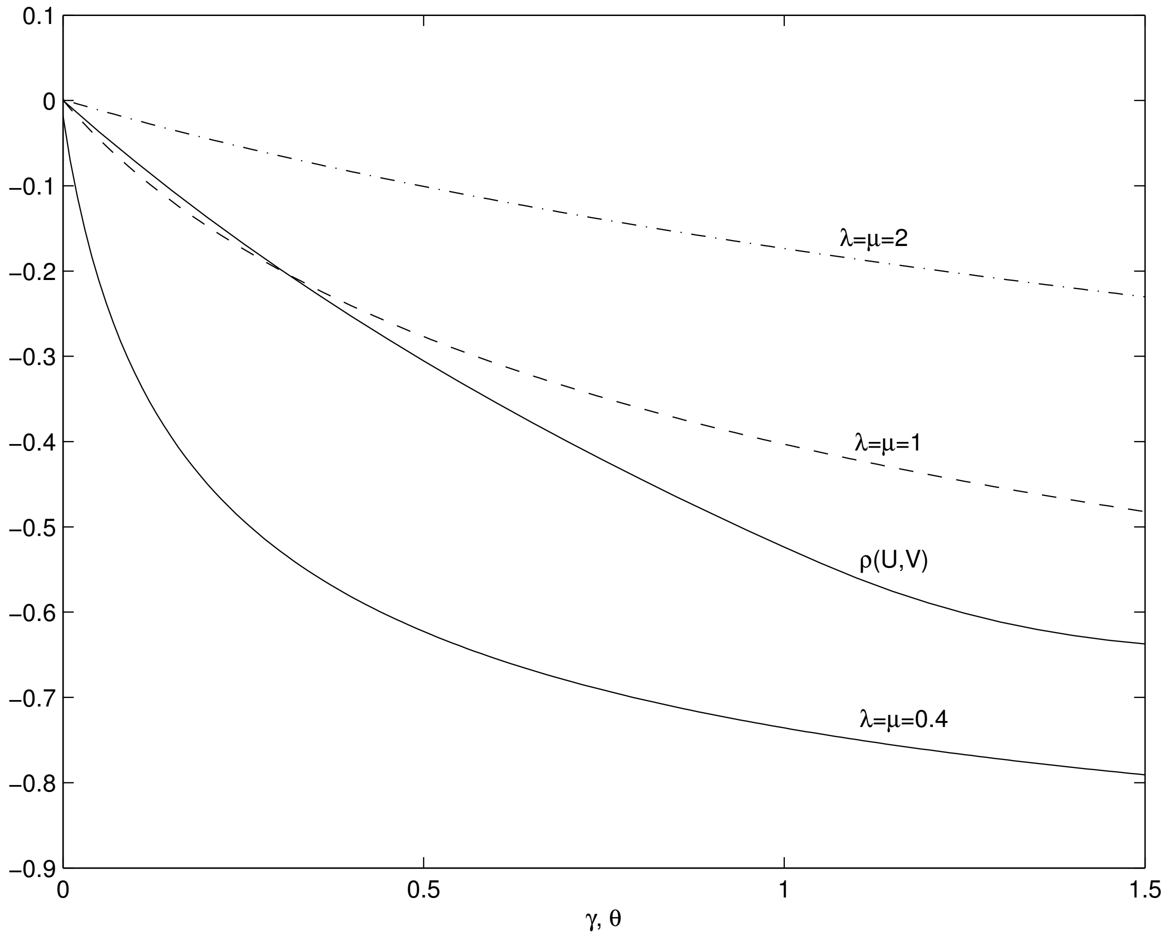

Figure 1: Dependence of $\rho(U, V)$ on parameter θ (solid curve) and $\rho(S, T)$ on γ, for given μ and λ, when $S \sim \exp(\lambda)$, $T \sim \exp(\mu)$.

Identifiability of the model consisting of the Barnett copula (2) and exponential marginal distributions has already been proved by Basu and Ghosh (1978). The model leads to the following likelihood function, given the data $\{z_i, \delta_i, \, i = 1, \ldots, N\}$:

$$L = \prod_{i=1}^{N} \left(\frac{-\partial \overline{F}_2(s,t)}{\partial s} \right)^{[\delta_i = 1]} \cdot \left(\frac{-\partial \overline{F}_2(s,t)}{\partial t} \right)^{[\delta_i = 2]} \cdot \overline{F}_2(s,t)^{[\delta_i = 0]} \Big|_{s=t=z_i} =$$

$$= \prod_{i=1}^{N} (\lambda + \gamma z_i)^{[\delta_i = 1]} \cdot (\mu + \gamma z_i)^{[\delta_i = 2]} \cdot \overline{F}_2(z_i, z_i). \tag{3}$$

3. Modelling the time to first goal

We shall use the data from the Czech First League, season 2014–15. Sixteen participating teams played together 240 matches (i.e. twice with each other, home and away). 21 matches ended without goals. More information on the Czech football league can be found on `http://www.sport.cz/fotbal/synot-liga/#vysledky`. Some statistics is provided also in Volf (2016).

As in the standard model of Maher (1982), each team (i) is characterized by its attack parameter a_i and defence parameter b_i, parameter h denotes the advantage of home field. The scoring intensities to the first goal in a match between home team i and away team j are then given as $a_i \cdot b_j \cdot h$, $a_j \cdot b_i$, respectively. Consequently, the time to the 1-st goal arises from two competing exponential random variables

$$S_{ij} \sim \exp(a_i \cdot b_j \cdot h)\,,\; T_{ij} \sim \exp(a_j \cdot b_i).$$

However, only the incidence of the first of them is observed. Or, in the case of no-score draw (0:0), times are censored by a fixed value, $C_i \equiv 90$ minutes. Further, it was assumed that their mutual dependence can be expressed via the Barnett copula described in Part 2.3.

3.1. Results of analysis

We solved the problem of the maximum likelihood estimation (MLE) of 34 parameters: a_i, b_i of 16 teams, home advantage parameter h, and γ characterizing the dependence. It was assumed that both h and γ were the same in all matches. The results of the MLE of teams parameters are displayed in Table 1. For computational convenience, we estimated $\alpha_i = \ln a_i$, $\beta_i = \ln b_i$, also $\delta = \ln h$. The ML estimates of two common parameters (with half-widths of 95% confidence intervals) were

$$\widehat{\delta} = 0.6417\,(0.2046),\; \widehat{h} = \exp(\hat{\delta}) = 1.8997, \quad \widehat{\gamma} = 0.945\,(0.078).$$

The correlation in each particular match depends on the teams parameters and on two common parameters h and γ. Its value can be traced roughly from Figure 1, or computed from the corresponding two-dimensional exponential model. For instance in matches of two leaders, Plzen and Sparta, numerical computation yielded $\rho(S,T) = -0.569$. In a match of teams with rather poor attack and yet fair defence, as for instance for Bohemians and Jihlava, we obtained $\rho(S,T) = -0.800$ which could be interpreted that the first goal was even more important. Further, the value of the parameter $h = 1.9$ indicated that the chance that the home team scores first was about

Team	α		β		a	b
Plzen	0.9742	(0.4664)	−1.8874	(1.7569)	2.6490	0.1515
Sparta	0.3662	(0.6055)	−0.9755	(0.8235)	1.4422	0.3770
Jablonec	0.2115	(0.5584)	−1.4791	(1.1286)	1.2356	0.2278
Ml. Boleslav	0.8080	(0.5539)	−0.2759	(0.6479)	2.2433	0.7589
Pribram	−0.0464	(0.6898)	−0.7362	(0.7491)	0.9547	0.4789
Dukla	−0.2479	(0.8046)	−0.0606	(0.5797)	0.7804	0.9412
Teplice	0.0205	(0.6216)	−1.5465	(1.2794)	1.0207	0.2130
Bohemians	−1.3719	(1.4189)	−0.6103	(0.6467)	0.2536	0.5432
Slovacko	0.2151	(0.6469)	−0.2541	(0.6111)	1.2400	0.7756
Jihlava	−0.2056	(0.7296)	−0.8168	(0.7615)	0.8141	0.4419
Slavia	0.3320	(0.5780)	−0.7249	(0.7983)	1.3938	0.4843
Liberec	−0.5043	(0.8504)	−0.3311	(0.6083)	0.6039	0.7181
Ostrava	−0.3343	(0.7779)	−0.4247	(0.6289)	0.7159	0.6540
Brno	−0.6091	(0.9150)	0.0883	(0.5231)	0.5438	1.0923
Hradec Kr.	−0.3694	(0.8226)	−0.1606	(0.5757)	0.6912	0.8517
C. Budejovice	−0.0435	(0.9000)	0.3128	(0.4921)	0.9574	1.3672

Table 1: **Results:** Estimated parameters $\alpha_i = \ln a_i$, $\beta_i = \ln b_i$ (with half-widths of approximate 95% conf. intervals in brackets), then a_i, b_i. Values of parameters are related to time 90 minutes, in order to keep them in a reasonable scale (and to avoid numerical problems, too).

$1.9/2.9 = 0.66$, while in reality from 219 first goals, 129 were scored by home teams, $129/219 = 0.59$. Let us remark here that the estimated teams parameters are, to certain extent, just relative, that parameter values $a_i \cdot c$, b_i/c yield the same model, for any $c > 0$.

3.2. Brief statistics of the first goal impact

Let us provide also a brief statistics concerning the relationship between the first goal and the final result. From 240 matches of the season 21 ended without goals. From the remaining 219 matches there were 37 draws (other then 0:0), 116 home wins and 66 away wins. Further, from these 219 matches with goals, in 156 cases the team scoring first was also the winner, in 26 cases the opposite had occurred (and 37 ended by draw). Estimated proportion $\widehat{p} = 156/219 = 0.7123$ is significantly larger than 0.5 and an approximate (asymptotic) 95% confidence interval for this proportion equals (0.6524, 0.7723). On the other hand, the chance to turn over the score after obtaining the first goal can be described by the estimated proportion $\widehat{q} = 26/219 = 0.1187$, which is still significantly larger than zero, yielding a 95% confidence interval (0.0759, 0.1616).

	Order 2014–15	a(14)	b(14)	a(15)	b(15)	Order 2015–16
1	Plzen	2.6490	0.1515	1.8218	0.2592	1
2	Sparta	1.4422	0.3770	1.0447	0.6478	2
3	Jablonec	1.2356	0.2278	0.7470	0.5481	7
4	Ml. Boleslav	2.2433	0.7589	1.5209	0.6408	4
5	Pribram	0.9547	0.4789	0.6197	0.9066	14
6	Dukla	0.7804	0.9412	1.2129	0.5980	10
7	Teplice	1.0207	0.2130	0.8845	0.8982	12
8	Bohemians	0.2536	0.5432	0.8931	0.5389	9
9	Slovacko	1.2400	0.7756	0.3463	1.0957	8
10	Jihlava	0.8141	0.4419	0.7611	0.5699	11
11	Slavia	1.3938	0.4843	1.4646	0.3974	5
12	Liberec	0.6039	0.7181	1.0913	0.5251	3
13	Ostrava	0.7159	0.6540	1.3502	1.1322	16
14	Brno	0.5438	1.0923	0.6325	0.5334	6
15	H. Kr./Olomouc	0.6912	0.8517	0.4542	0.4836	15
16	C. Budej./Zlin	0.9574	1.3672	1.1406	0.9276	13

Table 2: Comparison of results, i.e. estimated parameters and final order of teams in seasons 2014–15 and 2015–16.

3.3. Comparison with season 2015–16

The purpose of this section is to examine how stable are the parameters during a longer period. The data from the Czech First League season 2015–16 have been analyzed in the same manner as the season previous. Table 2 displays estimated teams parameters (with two new teams). It is seen that, at least for 5 teams, their final order as well as their parameters differ significantly from the values in Table 1. Hence, it has no much sense to use both data-sets jointly, we shall obtain some average values with poor interpretation. As regards two common parameters, we have obtained

$$\widehat{\delta} = 0.4837\,(0.1966),\;\; \widehat{h} = \exp(\hat{\delta}) = 1.6221,\quad \widehat{\gamma} = 1.450\,(0.117).$$

They are not far from the former values, nevertheless, from the statistical point of view, estimated γ differs significantly (corresponding confidence intervals are disjoint). We guess that the results reflect a specific feature of the season 2015–16: it was rather rich to scored goals: 676 goals were scored, while in the preceding season their number was 645. Consequently, a smaller correlations could be expected. Regarding the correlation of exponential distributions in the framework of the Barnett copula model, it is influenced jointly by teams parameters, by h, and by γ. Then, though the γ is larger, the other parameters may make the correlations smaller (compare also with figure 1). For instance, now the match of Sparta and Plzen leads to estimated $\rho = -0.602$, for Bohemians against Jihlava the value $\rho = -0.676$ has been obtained. Notice that the first-goal intensity of the Bohemians has increased considerably.

Let us again summarize the first goals statistics. From 240 matches of the 2015–16 season just 12 ended without goals. From remaining 228 matches there were 46 draws (other then 0:0), 119 home and 63 away wins. The home team has scored first in 134 cases, the away team in 94 cases. The proportion characterizing the home advantage equals then $134/228 = 0.59$, while the home advantage parameter leads to $\widehat{h}/(1+\widehat{h}) = 1.62/2.62 = 0.62$, which is a comparable value. Further, from 228 matches with goals, in 146 cases the team scoring first was also the winner, in 36 cases the the first scoring team has lost. Estimated proportions are now $\widehat{p} = 146/228 = 0.6404$ with asymptotic 95% confidence interval (0.5785, 0.7026), and $\widehat{q} = 36/228 = 0.1579$ (0.1106, 0.2052). Both are comparable with proportions from the season 2014–15.

4. Discussion and concluding remarks

The results lead to a conclusion that the correlation is, as a rule, negative, and is absolutely larger in more competitive matches, i.e. the matches of teams with good defence and comparable attack abilities. This conclusion is thus comparable with the results of McHale and Scarf (2011). The analysis has been repeated with the use of Gauss copula (Volf, 2016). Here, the correlation was common for all matches, i.e. not so flexible. Nevertheless, its estimated value was negative, too, and teams parameters comparable with those reported here. It has to be pointed out that the parameters estimated above concern just the stage of a match up to the first goal. It can be expected that the team performance changes during the match and is related to the actual score, elapsing time, and to other factors characterizing the match state. This aspect is also reflected by more advanced models of score development; see again for instance Dixon and Robinson (1998), Volf (2009) and an overview of models provided there. Hence, the approach proposed in the present study can be extended to the analysis of times to next goals. Another generalization can consider different copula parameters for certain groups of matches or teams.

In spite of the fact that the identification of proposed model, and therefore also the consistency of parameter estimates, are guaranteed theoretically, simulated experiments show rather slow convergence of estimates to 'true' values. It is also seen that the confidence intervals for parameters (approximate, i.e. based on asymptotic normality of estimates) are rather wide, which is a natural consequence of a rather high ratio of the number of parameters to the number of matches. Simultaneously, however, this follows from the fact that the log-likelihood function, as a function of γ, is rather flat.

References

[1] Basu A. P., Ghosh J. K. (1978): Identifiability of the Multinormal and Other Distributions under Competing Risks Model. *Journal of Multivariate Analysis* **8**, 413–429.

[2] Dixon M. J., Robinson M. E. (1998): A birth process model for association football matches. *The Statistician* **47**, 523–538.

[3] Escarela G., Carriere J. F. (2003): Fitting competing risks with an assumed copula. *Statistical Methods in Medical Research* **12**, 333–349.

[4] Kaishev V. K., Dimitrova D. S., and Haberman S. (2007): Modelling the joint distribution of competing risks survival times using copula functions. *Insurance: Mathematics and Economics* **41**, 339–361.

[5] Karlis D., Ntzoufras I. (2003): Analysis of sports data by using bivariate Poisson models. *J. R. Stat. Soc. Ser. D* **52**, 381–394.

[6] Lin D. Y. (1997): Non-parametric inference for cumulative incidence functions in competing risks studies. *Statistics in Medicine* **16**, 901–910.

[7] Maher M. J. (1982): Modelling association football scores. *Stat. Neerl.* **36**, 109–118.

[8] McHale I., Scarf P. A. (2011): Modelling the dependence of goals scored by opposing teams in international soccer matches. *Statistical Modelling* **11**, 219–236.

[9] Nevo D., Ritov Y. (2013): Around the goal: examining the effect of the first goal on the second goal in soccer using survival analysis methods. *Journal of Quantitative Analysis in Sports* **9**, 165–177.

[10] Schwarz M., Jongbloed G., and Van Keilegom I. (2013): On the identifiability of copulas in bivariate competing risks models. *Canadian Journal of Statistics* **41**, 291–303

[11] Tsiatis A. (1975): A nonidentifiability aspects of the problem of competing risks. *Proc. Nat. Acad. Sci. USA* **72**, 20–22.

[12] Volf P. (2009): A random point process model for the score in sport matches. *IMA Journal of Management Mathematics* **20**, 121–131.

[13] Volf P. (2016): On Statistical Analysis of Competing Risks with Application to the Time of First Goal. *Journal of Mathematics and Statistical Science* **2**, 606–623.

[14] Zheng M., Klein J. P. (1995): Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika* **82**, 127–138.

Obsah

Vědecké a odborné články

Zprávy a informace

Informační bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání. Příležitostně i mimořádné české a anglické číslo. Vydavatelem je Česká statistická společnost, IČ 00550795, adresa společnosti je Na padesátém 81, 100 82 Praha 10. Evidenční číslo registrace vedené Ministerstvem kultury ČR dle zákona č. 46/2000 Sb. je E 21214. Časopis je na Seznamu recenzovaných neimpaktovaných periodik vydávaných v ČR, více viz server `http://www.vyzkum.cz/`.

The Information Bulletin of the Czech Statistical Society is published quarterly.
The contributions in the journal are published in English, Czech and Slovak languages.

DOI: 10.5300/IB, http://dx.doi.org/10.5300/IB
ISSN 1210–8022 (Print), ISSN 1804–8617 (Online)

Toto číslo bylo vytištěno s laskavou podporou Českého statistického úřadu.