Liet. matem. rink., T. 42, spec. nr., 518–522

Konkuruojančių rizikų regresinis Farlie–Gumbel–Morgenstern skirstinių šeimos išgyvenamumo modelis

Audronė JAKAITIENĖ (VDU), Danas ZUOKAS (VDU)
el. paštas: iiauja@vdu.lt

Įvadas

Konkuruojančių rizikų teorija nagrinėja objektus, kurių gedimas (techninio elemento, sistemos atveju) ar mirtis (gyvo organizmo atveju) gali priklausyti nuo kelių priežasčių. Pastarosios vadinamos konkuruojančiomis rizikomis. Statistinė konkuruojančių rizikų teorija nagrinėja cenzūruotas imtis. Norima nustatyti, kokios rizikos labiausiai įtakoja gedimų (mirčių) intensyvumą ir laiką. Išgyvenamumo uždaviniuose konkuruojančios rizikos susijusios su populiacijos individų mirtingumo tyrimais, kai visi individai yra veikiami tų pačių mirties priežasčių (konkuruojančių rizikų). Pagrindinis tokių tyrimų tikslas yra išskirti duotos priežasties ar priežasčių komplekso poveikį populiacijai [2].

Išgyvenamumo funkcija

Tarkime, kad populiaciją vienu metu veikia m mirties priežasčių $C_1, \ldots, C_m$. Kiekvieną mirties priežastį atitinka hipotetinis (potencialus) mirties laikas $T_1 \geqslant 0, \ldots, T_m \geqslant 0$. Išgyvenamumo funkcija vektoriui $(T_1, \ldots, T_m)'$ (daugiamatė išgyvenamumo funkcija) apibrėžiama taip:

$$S_{1\ldots m}(t_1, \ldots, t_m) = P\{T_1 > t_1, \ldots, T_m > t_m\}, \tag{1}$$

kur

$$S_j(t_j) = 1 - F_j(t_j) = S_{1\ldots m}(0, \ldots, t_j, \ldots, 0) \tag{2}$$

yra marginalinės išgyvenamumo funkcijos, o $F_j(t_j)$ – marginalinė pasiskirstymo funkcija. Čia ir toliau $j = 1, 2, \ldots, m$.

Konkretus stebimas mirties laikas yra $T = \min(T_1, \ldots, T_m)$. Todėl bendra išgyvenamumo funkcija apibrėžiama taip (yra vieno kintamojo funkcija):

$$S_T(t) = P\{T > t\} = S_{1\ldots m}(t, \ldots, t). \tag{3}$$

Taip apibrėžtos bendros išgyvenamumo funkcijos bendra (overall) rizikos funkcija yra

$$h_T(t) = -\frac{\mathrm{d}\log S_T(t)}{\mathrm{d}t}, \tag{4}$$

o atskiros (crude) rizikos funkcijos yra [4]

$$h_j(t_1,\ldots,t_m) = -\frac{\partial \log S_{1\ldots m}(t_1,\ldots,t_m)}{\partial t_j}\Big|_{t_1=\ldots=t_m=t}. \tag{5}$$

FGM skirstinių šeimos išgyvenamumo funkcija

Šiame darbe daugiamate išgyvenamumo funkcija buvo pasirinkta trimatė Farlie–Gumbel–Morgenstern (FGM) skirstinių šeimos išgyvenamumo funkcija:

$$\begin{aligned} S_{123}(t_1,t_2,t_3) = {} & S_1(t_1)\,S_2(t_2)\,S_3(t_3)\left[1 + a_{12}F_1(t_1)\,F_2(t_2)\right. \\ & \left. + a_{13}F_1(t_1)\,F_3(t_3) + a_{23}F_2(t_2)\,F_3(t_3)\right]. \end{aligned} \tag{6}$$

a_{12}, a_{13} ir a_{23} yra priklausomumo parametrai tarp marginaliniais skirstiniais aprašytų atsitiktinių dydžių. Jie tenkina keturių nelygybių sistemą [3]:

$$\begin{cases} 1 + a_{12} + a_{13} + a_{23} \geqslant 0, \\ 1 + a_{12} - a_{13} - a_{23} \geqslant 0, \\ 1 - a_{12} + a_{13} - a_{23} \geqslant 0, \\ 1 - a_{12} - a_{13} + a_{23} \geqslant 0. \end{cases} \tag{7}$$

Atsitiktinių dydžių T_1, T_2, T_3 marginaliniais skirstiniais buvo pasirinkti nupjautieji iš kairės (apibrėžti neneigiamiems atsitiktiniams dydžiams $T_1 \geqslant 0$, $T_2 \geqslant 0$, $T_3 \geqslant 0$) logistiniai skirstiniai, kurių išgyvenamumo funkcija yra:

$$S_j(t_j, Z) = \frac{\exp\left\{-\frac{t_j}{\beta_j Z'}\right\}\left(1 + \exp\left\{\frac{\alpha_j Z'}{\beta_j Z'}\right\}\right)}{1 + \exp\left\{-\frac{t_j - \alpha_j Z'}{\beta_j Z'}\right\}}, \tag{8}$$

$\alpha_j Z' = (\alpha_{j0}\ \ldots\ \alpha_{jk})(1\ Z_1\ \ldots\ Z_k)' = \alpha_{j0} + \alpha_{j1}Z_1 + \cdots + \alpha_{jk}Z_k \geqslant 0$ yra tiesinė regresorių, nusakančių atsitiktinio dydžio vietą, kombinacija, o $\beta_j Z' = (\beta_{j0}\ \ldots\ \beta_{jk})(1\ Z_1\ \ldots\ Z_k)' = \beta_{j0} + \beta_{j1}Z_1 + \cdots + \beta_{jk}Z_k > 0$ – tiesinė regresorių, nusakančių atsitiktinio dydžio mastelį, kombinacija. Todėl sakoma, kad išgyvenamumo modelis yra regresinis.

Modelio parametrų įvertinimas

Nežinomų modelio parametrų (marginalinių funkcijų vietos ir mastelio, ir ryšio tarp atsitiktinių dydžių parametrus) įvertinimui taikomas didžiausio tikėtinumo metodas. Įveskime cenzūravimo indikatorių:

$$\delta_j = \begin{cases} 1, & \text{jei } T \leqslant T^+, \text{ ir } T = T_j, \text{ t.y., mirties priežastis yra } j,\ X = T_j, \\ 0, & \text{jei stebėjimas yra cenzūruotas, } X = T^+, \end{cases} \tag{9}$$

čia T^+ yra cenzūravimo laikas. Taigi stebima N tūrio imtis, o kiekvienas imties objektas turi stebėjimo duomenis $(X_i, \delta_{j_i}, j_i, Z_i)$, $X_i = \min(T_i, T_i^+)$ – mirties arba cenzūravimo laikas, δ_{j_i} cenzūravimo indikatorius, j_i – mirties priežasties indeksas ir Z_i – regresorių vektorius ($j_i = 1, 2, \ldots, m$, $i = 1, 2, \ldots, N$). Naudojant išgyvenamumo ir rizikos funkcijas sukonstruojama tikėtinumo funkcija:

$$\begin{aligned} &L(X_1, \ldots, X_N, \delta_{j_1}, \ldots, \delta_{j_N}, Z_i, \theta) \\ &\quad = \prod_{i=1}^{N} \left(h_{j_i}(X_i, Z_i, \theta)\right)^{\delta_{j_i}} S_T(X_i, Z_i, \theta). \end{aligned} \tag{10}$$

Darbe pasiūlytas ir kitas tikėtinumo funkcijos konstravimo būdas:

$$\begin{aligned} &L(X_1, \ldots, X_N, \delta_{j_1}, \ldots, \delta_{j_N}, Z_i, \theta) \\ &\quad = \prod_{i=1}^{N} \left(f_{j_i}(X_i, Z_i, \theta)\overline{S}_{j_i}(X_i, Z_i, \theta)\right)^{\delta_{j_i}} S_T(X_i, Z_i, \theta)^{1-\delta_{j_i}}, \end{aligned} \tag{11}$$

kur $f_j(X_i, Z_i, \theta)$ – (2) formule išreikštą išgyvenamumo funkciją atitinkantis vienmatis marginalinis j-osios mirties priežasties potencialaus mirties laiko pasiskirstymo tankis, apskaičiuotas i-ojo stebėjimo laiku X_i. θ – vertinamų parametrų vektorius (priklausomumo (7), padėties ir mastelio (8)). $\overline{S}_{j_i}(X_i, Z_i, \theta)$, pavadinkime ją "sąlygine" išgyvenamumo funkcija, apibrėžiama taip:

$$\begin{aligned} \overline{S}_j(t_j) &= P\{T_k > t_k, k \in \{1, \ldots, m\}\backslash\{j\} \mid T_j \leqslant t_j\} \\ &= \frac{\int_{t_j}^{\infty}\int_{t_j}^{\infty}\int_0^{t_j} f_{1\ldots m}(u_1, \ldots, u_m)\,\mathrm{d}u_k\mathrm{d}u_j}{\int_0^{t_j} f_j(u_j)\,\mathrm{d}u_j}, \quad (k \in \{1, \ldots, m\}\backslash\{j\}). \end{aligned} \tag{12}$$

Čia $f_{1\ldots m}(t_1, \ldots, t_m)$ yra daugiamatis, (1) formule išreikštą daugiamatę pasiskirstymo funkciją atitinkantis, potencialių mirties laikų vektoriaus pasiskirstymo tankis. Tokio tikėtinumo funkcijos konstravimo būdo prasmė yra ta, kad atsižvelgiama ne tik į tai, kad objektas mirė nuo j-osios priežasties, bet ir į tai, kad jis nemirė nuo likusių priežasčių, t.y., yra cenzūruotas stebėjimas kitų mirties priežasčių atžvilgiu.

Praktinis modelio pritaikymas

Tikėtinumo funkcijos sprendinys ieškotas skaitmeniškai, naudojant BFGS optimizavimo algoritmą su apribojimais [1]. Yra žinoma, kad didžiausio tikėtinumo metodu rastas įvertis, esant atsitiktiniam cenzūravimui, asimptotiškai pasiskirstęs pagal daugiamatį normalųjį dėsnį su vidurkiu θ (vertinamų parametrų vektorius) ir kovariacine matrica $I^{-1}(\theta)$, t.y.,

$$\widehat{\theta} \sim N(\theta, I^{-1}(\theta)) \quad \theta \in \mathbb{R}^p. \tag{13}$$

I^{-1}, Fišerio informacinės matricos atvirkštinė, maksimumo taške nusako kovariaciją tarp maksimalaus tikėtinumo įverčio $\widehat{\theta}$ komponenčių. Tikrinant hipotezę apie parametrus H_0: $\widehat{\theta} = \theta_0$, buvo naudojamas Neyman–Pearson–Wilks kriterijus

$$k = -2 \log \frac{L(\theta_0)}{L(\widehat{\theta})}, \tag{14}$$

ir yra žinoma, kad jis turi χ^2_p asimptotinį pasiskirstymą [4].

Rezultatai

Šiame darbe buvo sudarytas trimatis regresinis konkuruojančių rizikų išgyvenamumo modelis FGM skirstinių šeimai su logistiniais iš kairės nupjautaisiais marginaliniais skirstiniais. Pasiūlytas tikėtinumo funkcijos konstravimo būdas išgyvenamumo modelio parametrams įvertinti. Sudarytas išgyvenamumo modelis pritaikytas 1972 metais pradėtos

1 lentelė

Stebėtų mirčių bei pagal modelį dviem variantais ((10) ir (11)) paskaičiuotų mirčių skaičius

Mirties priežastis	Koronarinės ligos	Vėžys	Kitos	Viso
Pirmas variantas, (10) formulė				
Stebėtų mirčių skaičius	412	247	201	860
Modeliavimo rezultatai	533	353	239	1125
Skirtumas	121	106	38	265
% nuo stebėtų mirčių	29	43	19	31
Antras variantas, (11) formulė				
Stebėtų mirčių skaičius	412	247	201	860
Modeliavimo rezultatai	469	320	245	1034
Skirtumas	57	73	44	174
% nuo stebėtų mirčių	14	30	22	20

ir 22 metus trukusios Pasaulio sveikatos organizacijos koordinuotos, Kauno–Roterdamo profilaktinės metodinės programos, pavadintos "KRIS" (The Kaunas Rotterdam Intervention Study), metu surinktiems duomenims.

1 lentelėje yra pateikti stebėtų mirčių nuo kiekvienos priežasties bei pagal modelį paskaičiuotų mirčių nuo kiekvienos priežasties skaičiai abiem variantais ((10) ir (11)). Modeliavimo mirčių skaičius rastas pagal tokią formulę:

$$\mathrm{MS}_j = \sum_{i=1}^{N} F_{j_i}(X_i)\,\mathbf{1}\{j_i = j\}, \quad j = 1,2,3, \tag{15}$$

t.y., buvo susumuotos visų stebėjimų tikimybės mirti nuo konkrečios priežasties per visą stebėjimo laikotarpį. Tokiu būdu gautas tam tikras skaičius, nusakantis kiekvienos mirties dažnį.

Iš 1 lentelėje pateiktų rezultatų matome, kad pagal pasiūlytą variantą (įvedant "sąlyginę" išgyvenamumo funkciją (11)) sukonstruotos tikėtinumo funkcijos maksimizavimas duoda geresnius rezultatus nei pagal literatūroje rastą variantą (naudojant išgyvenamumo ir rizikos funkcijas (10)) sukonstruotos tikėtinumo funkcijos maksimizavimas, ta prasme, kad pagal modelį paskaičiuotų mirčių skaičius pirmu atveju buvo artimesnis stebėtų mirčių skaičiui nuo koronarinių ligų ir vėžio, ir šiek tiek didesnis mirtims nuo kitų priežasčių.

Literatūra

[1] T. Coleman, M.A. Branch, A. Grace, *Optimization Toolbox User's Guide*, MathWorks (1990–1999).
[2] A. Jakaitienė, *Konkuruojančių rizikų regresinių modelių algoritmai*, Daktaro disertacija, Kaunas (2001).
[3] S. Kotz, N. Balakrishnan, N.L. Johnson, *Continuous Multivariate Distributions: Models and Applications*, Vol. 1, John Willey and Sons, New York (2000).
[4] R.G. Miller, Jr., *Survival Analysis*, John Willey and Sons, New York (1981).

The survival regression model of competing risks for the family of Farlie–Gumbel–Morgenstern distributions

A. Jakaitienė, D. Zuokas

In this paper the trivariate survival regression model for FGM family of distributions is constructed with marginal left-truncated logistic distributions. Two methods (using survival and hazard functions in the first case, and distributional density and "conditional" survival function in the second case) are used when constructing likelihood function for model parameter estimation. Constructed survival model was run with the data of the "KRIS" (The Kaunas Rotterdam Intervention Study), which lasted for 22 years from 1972. The results show, that using second case for likelihood function construction gives better approximation for the data, because some additional conditions are considered.