A RIZIKÓFÜGGVÉNY LOGISZTIKUS MODELLJE EPIDEMIOLÓGIAI VIZSGÁLATOKBAN

Békéssy András, Krámli András, Soltész János, Csukás Andrásné
MTA SZTAKI, Országos Kardiológiai Intézet

Előadásunknak az a célja, hogy néhány uj szemponttal járuljon hozzá a logisztikus függvénnyel történő rizikóbecsléssel kapcsolatban összegyült tapasztalatokhoz. A logisztikus modell egyike az epidemiológiában leggyakrabban használt modelleknek. Az epidemiológia egyik alapfeladata az, hogy nagy populációban bizonyos könnyen mérhető adatok /un. rizikófaktorok/ alapján minden egyedre előrejelezzék egy adott betegség fellépésének valószinüségét. A rizikófaktorok általában nem állnak közvetlen kapcsolatban a szóbanforgó betegséggel, ezért az egyik legnehezebb feladat a lényeges rizikófaktorok kiválasztása. Matematikai szempontból a fenti előrejelzési feladat az un. statisztikai diszkriminancia-analizissel rokon. A diszkriminancia-analizis alapfeladata a következő: Adva van n populáció /$A_1,\ldots,A_n$/ és minden egyes populációból bizonyos előre rögzitett elemszámu minta. A mintához tartozó egyedek bizonyos folytonos illetve diszkrét jellemzői alapján megszerkesztünk egy olyan szabályt, amelynek segitségével további egyedekről eldönthetjük, hogy melyik osztályba tartoznak.

Az epidemiológiában felmerülő szétválasztási probléma különlegessége az, hogy az egyik populáció /egy adott időszakban megbetegedett személyek/ viszonylag kicsi. A megelőzés szempontjából az a legfontosabb feladat, hogy a várhatóan megbetegedő személyeknek minél nagyobb hányadát lehetőleg kevés paramétertől függő szabály segitségével válasszuk ki. Ezt a szétválasztási szabályt egy olyan minta alapján szerkesztjük meg, amelyben az egyedek már osztályozva vannak. Először minden populációban - paraméteres vagy nem paraméteres eljárással - megbecsüljük az $x_1,\ldots,x_k$ rizikóváltozók eloszlásfüggvényét, majd Bayes tétele segitségével minden j egyedre kiszámitjuk annak a $\Pi_i(x_{j1},\ldots,x_{jk})$ a-posteriori valószinüségét, hogy az A_i osztályba tartozik, ha

$$\max_l \sum_m \Pi_m(x_{j1},\ldots,x_{jk})\, r_{lm} = \sum_m \Pi_m(x_{j1},\ldots,x_{jk})\, r_{im}$$

ahol $r_{l,m}$ annak a hibás osztályozásnak a költsége, ha egy A_l osztályba tartozó egyedet az A_m osztályba sorolunk.

Ha csak két osztály van (A_1, A_2), és a rizikóváltozók közös Σ kovarianciamátrixu, különböző /csak az osztálytól függő/ várhatóértékü együttesen normális eloszlásu változók, akkor az osztályozás feladata leegyszerüsödik: a j-edik egyedet akkor soroljuk pl. az A_2 osztályba, ha a $\Pi_2(x_{j1},\ldots,x_{jk})$ a-posteriori valószinüség egy bizonyos szintnél nagyobb. Ekkor $\Pi_2(x_{j1},\ldots,x_{jk})$ a jól ismert logisztikus függvény

$$\Pi_2(x_1,\ldots,x_k) = \left[1 + \exp(-\alpha - \sum_{l=1}^{k} \beta_l x_l)\right]^{-1}$$

$$\beta_l = \sum_{s=1}^{k} (m_{2s}-m_{1s})\sigma^{ls} \qquad (1)$$

ahol $m_{11},\ldots,m_{1k}$ illetve $m_{21},\ldots,m_{2k}$ az A_1 ill. A_2 osztályban a rizikóváltozók várhatóértéke, $\{\sigma^{ls}\}$ a Σ kovarianciamátrix inverze, N_1 illetve N_2 az A_1 illetve A_2 osztályból vett minták elemeinek száma, végül

$$\alpha = -\frac{1}{2}\sum_{s=1}^{k} \beta_s(m_{1s}+m_{2s}) - \log\frac{N_1}{N_2} \quad .$$

Σ, m_1, m_2 a minta alapján becsülhető.

Még ha el is fogadjuk azt a feltevést, hogy a rizikófaktorok együttesen normális eloszlásuak, lehetséges, hogy a különböző populációkhoz tartozó kovarianciamátrixok nem azonosak. Ebben az esetben a megbetegedés két rizikófaktor együttes hatásával kapcsolatos. Ekkor a Bayes-féle rizikófüggvény a

$$\Pi_2(x_1,\ldots,x_k)=\left\{1+\frac{N_1}{N_2}\frac{\det\Sigma_2}{\det\Sigma_1}\exp\left[-\frac{1}{2}\sum_{l,s}\sigma_1^{ls}(x_l-m_{1l})(x_s-m_{1s})+\right.\right.$$
$$\left.\left.+\frac{1}{2}\sum_{l,s}\sigma_2^{ls}(x_l-m_{2l})(x_s-m_{2s})\right]\right\}^{-1} \qquad (2)$$

alakot ölti, ahol $\{\sigma_1^{ls}\}$ ill. $\{\sigma_2^{ls}\}$ az A_1 ill. A_2 populációhoz tartozó kovarianciamátrixok inverzei.

Az (1) esetben a rizikófüggvény szintfelületei hipersikok, a (2) esetben pedig kvadratikus felületek.

Példaként tekintsük az ischaemiás szivbetegségek rizikójának becslésére három Szeged környéki községben végzett vizsgálatokat /ezek hasonlóak ahhoz a hét országban végzett vizsgálatokhoz, amelyet Ancel Keys szervezett/. Először elfogadtuk mind az együttes normalitás, mind pedig a kovarianciamátrixok azonosságának feltételét. A numerikus eredmények közül a következő táblázatot választottuk ki illusztrációként:

Az x_1,x_2,x_3,x_4 rizikóváltozó rendre a systoles vérnyomás, a koleszterin, az életkor és a Quetelet index.
Ekkor

α=-15.272, β_1=0.013, β_2=0.006, β_3=0.091 β_4=0.159

Illeszkedési tábla

decilis	1	2	3	4	5	6	7	8	9	10
egyedek száma	101	101	101	101	100	100	101	101	101	101
várható megbetegedésszám	0.50	0.77	1.07	1.38	1.73	2.18	2.87	3.89	5.77	13.14
megfigyelt megbetegedésszám	0	1	1	3	3	4	1	2	6	11

Pontosabb analizis azt mutatta, hogy a kovarianciamátrixok szignifikánsan eltérnek a két populációban. A (2) rizikófüggvényt használva ugyanazokra a rizikófaktorokra a következő illeszkedési táblát kaptuk.

decilis	1	2	3	4	5	6	7	8	9	10
egyedek száma	101	101	101	101	100	100	101	101	101	101
várható megbet.sz.	0.25	0.51	0.78	1.16	1.64	2.20	2.91	3.79	5.09	18.63
megf. megbet.sz.	1	1	2	0	1	1	4	5	5	12

Az első esetben 20, a másodikban pedig 26 /sulyosan/ beteg esett a felső négy decilisbe. Igy a második formula alkalmasabb azoknak az egyedeknek kiválasztására, akiknek az ischaemiás szivbetegségre való hajlamuk viszonylag nagy. A (2) formula azonban sokkal több paramétert tartalmaz, igy ez a modell kevésbé stabil.

A logisztikus modellt egyszerüsége miatt még akkor is gyakran használják, amikor a rizikóváltozók együttes normalitásának feltétele nem fogadható el. Walker és Duncan [1] a modell paramétereinek becslésére egy iterativ eljárást ajánl. Módszerük szoros kapcsolatban áll a számitandó α, β_1, ..., β_k paraméterek maximum likelihood becslésével, ha feltételezzük, hogy (1) alaku annak az eseménynek a valószinüsége, hogy egy személy az A_2 osztályba tartozik /azaz beteg/. [2]-ben megmutattuk, hogy Walker és Duncan formulái azonosak a likelihood egyenlet Taylor-sorfejtésének elsőrendü tagjaival.

Woodbory, Manton és Stallard [3] krónikus betegségek longitudinális vizsgálatánál több logisztikus függvény szorzatát ajánlották. Ehhez hasonló eredményre jutunk, ha feltételezzük, hogy a rizikófüggvény szintfelületei két vagy több hipersikból állnak. A Bayes-módszer is ilyen rizikófüggvényhez vezet, ha a nagyobb /egészséges/ populáció alpopulációkra osztható, feltételezve az együttes normalitást és a közös kovarianciamátrixot. Vizsgálatunkban két alpopulációt természetes módon meg lehet adni: volt 69 személy, aki kevésbé sulyos ischaemiás szivbetegségben szenvedett. Ebben az esetben az A_2 /sulyos beteg/ populációhoz tartozó rizikófüggvény a

$$\Pi_2(x_1,\ldots,x_k)=\frac{N_2^* \exp\left\{-\frac{1}{2}\sum_{l,s=1}^{k}\sigma^{ls}(x_l-m_{2l})(x_s-m_{2s})\right\}}{\sum_{i=1}^{3}\left[N_i \exp\left\{-\frac{1}{2}\sum_{l,s=1}^{k}\sigma^{ls}(x_l-m_{il})(x_s-m_{is})\right\}\right]} \qquad (3)$$

alakot ölti, ahol m_1, m_2 és m_3 rendre az A_1, A_2 és A_3 /egészséges, sulyos beteg, kevésbé sulyos beteg/ populáció rizikóváltozói várható értékeinek vektora. Ugyanazon rizikófaktorok esetén a következő illeszkedési táblát kaptuk:

decilis	1	2	3	4	5	6	7	8	9	10
egyedek száma	101	101	101	101	100	100	101	101	101	101
várható megbet.sz.	0.48	0.74	1.06	1.37	1.73	2.19	2.89	3.93	5.81	13.11
megf. megbet.szám	0	1	2	2	3	4	1	2	6	11

Ez utóbbi tábla majdnem ugyanaz mint az első, mivel az m_1, m_2 és m_3 várhatóérték vektor jó közelitéssel egy egyenesre illeszkedett a mintatérben. Igy a mi esetünkben ez a bonyolultabb elválasztási szabály nem hatékonyabb a klasszikus logisztikus modellnél. Ez a megközelités azonban mégis igéretes, mivel ujabb szabadságfokokat ad anélkül, hogy lényegesen megnövelné a modell paramétereinek számát. Elképzelhető, hogy másképpen definiált alappopulációk esetén jobb eredményt kapnánk.

A Bayes-féle döntési eljárást alkalmaztuk arra az esetre is, amikor a rizikófaktorok között diszkrét változók /$d_1,\ldots,d_s$/ is előfordulnak. Feltételezzük, hogy a diszkrét változók teljesen függetlenek, és a folytonos vátlozóktól is függetlenek. Ekkor a megbetegedés a-posteriori valószinüsége

$$\Pi_2(x_1\ldots,x_k;d_1,\ldots,d_s) = \left\{1+\left[\prod_{u=1}^{s}\frac{p_1(d_u)}{p_2(d_u)}\right]\exp\left[-\alpha-\sum_{l=1}^{k}\beta_l x_l\right]\right\}^{-1} \qquad (4)$$

ahol $p_1(d_u)$/ill. $p_2(d_u)$/ annak az eseménynek a valószinüsége, hogy egy egészséges /ill. beteg/ személy u-adik diszkrét rizikóváltozója a d_u értéket veszi fel. Ezeket a valószinüségeket a relativ gyakoriságokkal becsülhetjük, az α, $\beta_1,\ldots,\beta_k$ paraméterekre pedig ugyanaz a becslés adható, mint az (1) esetben.

A systoles vérnyomást, a koleszterint, az életkort és a Quetelet indexet használva folytonos, az EKG V. Minnesota kódját és a dohányzás Keys-féle kódját használva a diszkrét rizikófaktorként, a (4) esetben a következő illeszkedési táblát kapjuk:

decilis	1	2	3	4	5	6	7	8	9	10
egyedek száma	100	100	100	100	99	99	100	100	100	100
várh. megbet.sz.	0.34	0.64	0.92	1.22	1.53	2.00	2.62	3.60	5.51	14.87
megf. megbet.sz.	0	0	1	2	1	3	3	4	7	11

Egy illeszkedési táblát epidemiológiai szempontból annál jobbnak tekinthetünk, minél több beteg esik a felső decilisekbe. Megállapithatjuk, hogy a mi esetünkben a fent vázolt modellek ebből a szempontból nem mutatnak lényeges eltérést, ezért a legkönnyebben mérhető és számitható modellt tudjuk javasolni.

Hivatkozások

[1] Walker, S. H. and Duncan, D.E.: Estimation of the probability of an event as a function of several independent variables, Biometrika, Vol. 54, pp. 167-179, 1967.

[2] Békéssy, A., Csukás, M., Krámli, A. and Soltész, J.: Iterative methods for determining risk factors, Proc. of Third Hungarian Biometric Conference, 1981, pp. 201-204.

[3] Woodbury, M.A., Manton, K.G. and Stallard, E.: Longitudinal Models for Chronic Disease Risk: an Evaluation of Logistic Multiple Regression and Alternatives, International Journal of Epidemiolgy, Vol. 10, pp. 187-197, 1981.