Marta Zalewska
Warszawski Uniwersytet Medyczny
Wojciech Niemiro
Uniwersytet Mikołaja Kopernika, Toruń
oraz Uniwersytet Warszawski

Model autologistyczny i jego zastosowanie do danych epidemiologicznych

Badania ECAP (Epidemiologia Chorób Alergicznych w Polsce) doprowadziły do stworzenia ogromnej bazy danych. Zebrano ankiety od blisko 23 tysięcy losowo wybranych osób, z czego ok. 30% przeszło badania lekarskie w kierunku diagnostyki chorób alergicznych. Baza danych obejmuje ponad 1200 zmiennych dla pojedynczej osoby. Większość stanowią zmienne binarne. Model autologistyczny wydaje się bardzo obiecującym narzędziem opisu współzależności tych zmiennych.

Rozważany model ze zmiennymi objaśniającymi ma następującą postać. Obserwujemy losowe, binarne zmienne „objaśniane" $x = (x_1, \ldots, x_d) \in \{0,1\}^d$ oraz zmienne „objaśniające" $u = (u_1, \ldots, u_p) \in \mathbb{R}^p$. Zakładamy, że rozkład prawdopodobieństwa ma postać:

$$p_\beta(x|u) := \frac{1}{Z(\beta,u)} \exp\Bigg(\underbrace{\sum_{i=1}^{d}\sum_{j=1}^{d} \beta_{i,j} x_i x_j}_{\text{auto}-\text{regresja}} + \underbrace{\sum_{i=1}^{d}\sum_{j=1}^{p} \beta_{i,d+j} x_i u_j}_{\text{regresja}} \Bigg).$$

Elementy macierzy $\beta = (\beta_{i,j})$ wymiaru $d \times (d+p)$ (przy tym $\beta_{i,j} = \beta_{j,i}$ dla $i,j = 1,\ldots,d$) są parametrami tego rozkładu i trzeba je estymować. Obecność nieznanej stałej normującej $Z(\beta,u)$ utrudnia, dla dużego d, obliczanie estymatorów największej wiarogodności. Stosuje się i bada różne metody estymacji, w tym maksimum pseudo-wiarogodności. Typowo model autologistyczny stosuje się w statystyce przestrzennej. Specyfika zastosowań do analizy danych zgromadzonych w bazie ECAP jest inna. Możemy zakładać, że dysponujemy sporą próbką losową $(x[1],u[1]),\ldots,(x[n],u[n])$ z powyżej napisanego rozkładu autologistycznego, co pozwala stosować przybliżenia asymptotyczne. W tym kontekście pojawiają się nowe możliwości obliczania estymatorów największej wiarogodności za pomocą metod Monte Carlo. Przytoczymy wyniki obliczeń dla danych z bazy ECAP oraz dla danych syntetycznych. Porównamy kilka metod estymacji. Omówimy też bayesowską imputację brakujących danych przy pomocy pewnej odmiany parametrycznego bootstrapu.