Lietuvos matematikos rinkinys
Lietuvos matematikų draugijos darbai, ser. B
53 t., 2012, 260–263

ISSN 0132-2818
www.mii.lt/LMR/

Daugiamatis mažų dažnių vertinimo algoritmas

Leonidas Sakalauskas, Ingrida Vaičiulytė

Vilniaus universitetas, Matematikos ir informatikos institutas
Akademijos g. 4, LT-08663 Vilnius
E. paštas: sakal@ktl.mii.lt, ingrida_vaiciulyte@yahoo.com

Santrauka. Šiame darbe nagrinėjamas empirinio Bajeso metodo taikymas kelių mažų dažnių įvertinimui. Modeliuojant kelių kiekybinių rodiklių – retų įvykių skaičių tikimybes, yra laikoma, kad tikimybės pasiskirsčiusios pagal Puasono dėsnį su skirtingais parametrais, kurie yra atsitiktiniai koreliuoti Gauso dydžiai. Nežinomų parametrų įverčiai gaunami didžiausio tikėtinumo metodu. Išvestos lygtys, kurias turi tenkinti modelio parametrų didžiausio tikėtinumo įverčiai.

Raktiniai žodžiai: Bajeso metodas, tikėtinumo metodas, Puasono–Gauso modelis.

Įvadas

Epidemiologiniuose tyrimuose yra aktuali kelių retų įvykių tikimybių įvertinimo problema. Tokiais įvykiais gali būti sergamumas arba mirtingumas nuo kokių nors ligų ir pan. Paprastas santykinis įvertis įvykių tikimybėms vertinti nelabai tinka:

$$P = \frac{Y}{N},$$

čia Y – įvykių skaičius, N – populiacijos dydis, nes populiacijų dydžiai gali labai svyruoti ir gautų įverčių patikimumas gali ženkliai skirtis. Todėl šioms tikimybėms vertinti yra pasiūlytas empirinis Bajeso metodas. Darbe pasinaudojama *logit* modeliu, kuriame įvedamas pagalbinis kintamasis α, išreiškiamas per nagrinėjamo įvykio tikimybę, ir apibrėžiamas formule

$$\alpha = \ln \frac{P}{1-P}, \tag{1}$$

čia P yra tikimybė, kad priklausomas kintamasis įgys reikšmę 1, $(1-P)$ – tikimybė, kad priklausomas kintamasis įgys reikšmę 0 [6]. Vieno įvykio tikimybės vertinimo algoritmas panaudojant empirinį Bajeso metodą buvo sudarytas P.C. Bradley ir A.L. Thomas (2000), D. Clayton ir J. Kaldor (1987), R.K. Tsutakava ir kt. (1985) [2, 3, 7]. O šiame darbe empirinis Bajeso metodas pritaikytas kelių retų įvykių tikimybėms.

1 Puasono–Gauso modelis

Tegul turime populiacijų aibę $\Lambda = (A_1, A_2, \ldots, A_K)$, sudarytą iš K populiacijų, čia kiekviena populiacija A_j turi N_j individų, $j = \overline{1, K}$. Tarkime, kad stebint populiacijas

gali įvykti M tam tikrų įvykių (susirgimo, mirties, ...). Tikslas yra įvertinti šių įvykių tikimybes P_j^m, kai Y_j^m yra stebėtas m-to įvykio pasirodymų skaičius, $j = \overline{1,K}$ ir $m = \overline{1,M}$.

Empiriniame Bajeso metode yra priimama, kad įvykių skaičiai Y_j^m populiacijose yra pasiskirstę pagal Puasono dėsnį su parametrais $\lambda_{j,m} = N_j \cdot P_j^m$, $1 \leqslant m \leqslant M$, t. y., su tankiu [2, 3, 7]:

$$f\big(Y_j^m, \lambda_{j,m}\big) = e^{-\lambda_{j,m}} \frac{(\lambda_{j,m})^{Y_j^m}}{(Y_j^m)!}, \quad j = 1, \ldots, K.$$

Puasono–Gauso modelyje priimama, kad įvykių tikimybių *logitai* populiacijose yra pasiskirstę pagal daugiamatį normalųjį skirstinį su parametrais μ, Σ [2], t. y., *logit* (1) skirstinio tankis

$$\mathrm{g}(\alpha, \mu, \Sigma) = \frac{\exp(-(\alpha-\mu)^T \Sigma^{-1}(\alpha-\mu))}{\sqrt{|\Sigma|} \cdot (2\pi)^{\frac{M}{2}}}.$$

Tikimybės P_j^m įvertis yra apskaičiuojamas aposterioriniu vidurkiu:

$$P_j^m = \frac{\int_{-\infty}^{+\infty} \frac{1}{1+e^{-\alpha_j^m}} \prod_{k=1}^{M} f\Big(Y_j^k, \frac{N_j}{1+e^{-\alpha_j^k}}\Big) \mathrm{g}(\alpha, \mu, \Sigma)\, d\alpha}{D_j(\mu, \Sigma)},$$

čia

$$D_j(\mu, \Sigma) = \int_{-\infty}^{+\infty} \prod_{k=1}^{M} f\left(Y_j^k, \frac{N_j}{1+e^{-\alpha_j^k}}\right) \mathrm{g}(\alpha, \mu, \Sigma)\, d\alpha \tag{2}$$

yra įvykių skaičiaus aposteriorinė tikimybė j-je populiacijoje, $j = \overline{1,K}$. Bajeso metodas dažnai taikomas statistikoje minimizuojant tam tikras funkcijas išreikštas per aposteriorinio tankio integralą. Taigi, nežinomi parametrai μ, Σ yra vertinami didžiausio tikėtinumo metodu [2, 7]. Galima parodyti, kad logaritminė tikėtinumo funkcija yra:

$$\begin{aligned} L(\mu, \Sigma) &= -\sum_{j=1}^{K} \ln\left(\int_{-\infty}^{+\infty} \prod_{k=1}^{M} f\left(Y_j^k, \frac{N_j}{1+e^{-\alpha_j^k}}\right) \mathrm{g}(\alpha, \mu, \Sigma)\, d\alpha\right) \\ &= -\sum_{j=1}^{K} \ln\big(D_j(\mu, \Sigma)\big), \end{aligned} \tag{3}$$

kurią minimizavus gaunami parametrų μ, Σ įverčiai.

2 Didžiausio tikėtinumo funkcijos išvestinės

Tikėtinumo funkcija (3) gali būti diferencijuojama daugelį kartų pagal parametrus μ, Σ. Šios funkcijos pirmos eilės išvestinės yra:

$$\frac{\partial L(\mu, \Sigma)}{\partial \mu} = \sum_{j=1}^{K} \frac{\int_{-\infty}^{+\infty} \Sigma^{-1}(\alpha-\mu) \prod_{k=1}^{M} f\Big(Y_j^k, \frac{N_j}{1+e^{-\alpha_j^k}}\Big) \mathrm{g}(\alpha, \mu, \Sigma)\, d\alpha}{D_j(\mu, \Sigma)}, \tag{4}$$

$$\frac{\partial L(\mu,\Sigma)}{\partial \Sigma} = \sum_{j=1}^{K} \frac{\int_{-\infty}^{+\infty}(\Sigma^{-1}-\Sigma^{-1}(\alpha-\mu)(\alpha-\mu)^T\Sigma^{-1})\prod_{k=1}^{M} f\left(Y_j^k,\frac{N_j}{1+e^{-\alpha_j^k}}\right)\mathsf{g}(\alpha,\mu,\Sigma)\,d\alpha}{D_j(\mu,\Sigma)}. \tag{5}$$

Gautas išvestines (4), (5) prilyginus nuliui, t. y.,

$$\frac{\partial L(\mu,\Sigma)}{\partial \mu} = 0,$$
$$\frac{\partial L(\mu,\Sigma)}{\partial \Sigma} = 0,$$

ir atlikus nesudėtingus veiksmus, galima gauti lygtis, kurias tenkina Puasono–Gauso modelio parametrų įverčiai

$$\hat{\mu} = \frac{1}{K}\sum_{j=1}^{K} \frac{\int_{-\infty}^{+\infty}\alpha\cdot\prod_{k=1}^{M} f\left(Y_j^k,\frac{N_j}{1+e^{-\alpha_j^k}}\right)\mathsf{g}(\alpha,\mu,\Sigma)\,d\alpha}{D_j(\mu,\Sigma)}, \tag{6}$$

$$\hat{\Sigma} = \frac{1}{K}\sum_{j=1}^{K} \frac{\int_{-\infty}^{+\infty}(\alpha-\mu)(\alpha-\mu)^T\prod_{k=1}^{M} f\left(Y_j^k,\frac{N_j}{1+e^{-\alpha_j^k}}\right)\mathsf{g}(\alpha,\mu,\Sigma)\,d\alpha}{D_j(\mu,\Sigma)}. \tag{7}$$

Didžiausio tikėtinumo įverčius $\hat{\mu}$, $\hat{\Sigma}$ galima apskaičiuoti iš lygčių (6), (7) kintamos metrikos metodu [4], išreiškus integralus Ermito–Gauso kvadratūrinėmis formulėmis [1]. Be to, didžiausio tikėtinumo funkcijos minimizavimą ir integravimą galima atlikti pasinaudojus matematine programine įranga MATHCAD, MAPLE ir pan.

3 Puasono–Gauso modelio parametrų įverčiai

Taip pat, norint rasti parametrų μ, Σ didžiausio tikėtinumo įverčius, (6), (7) lygtis galima spręsti „fiksuoto taško iteracijų“ metodu [5]:

$$\hat{\mu}_{t+1} = \frac{1}{K}\sum_{j=1}^{K} \frac{\int_{-\infty}^{+\infty}\alpha\cdot\prod_{k=1}^{M} f\left(Y_j^k,\frac{N_j}{1+e^{-\alpha_j^k}}\right)\mathsf{g}(\alpha,\mu_t,\Sigma_t)\,d\alpha}{D_j(\mu_t,\Sigma_t)}, \tag{8}$$

$$\hat{\Sigma}_{t+1} = \frac{1}{K}\sum_{j=1}^{K} \frac{\int_{-\infty}^{+\infty}(\alpha-\mu_t)(\alpha-\mu_t)^T\prod_{k=1}^{M} f\left(Y_j^k,\frac{N_j}{1+e^{-\alpha_j^k}}\right)\mathsf{g}(\alpha,\mu_t,\Sigma_t)\,d\alpha}{D_j(\mu_t,\Sigma_t)}, \tag{9}$$

čia integralus (2), (8), (9) galima apskaičiuoti pasinaudojus Ermito-Gauso kvadratūrinėmis formulėmis [1].

Lygybėse (8)–(9) galima parinkti tokį pradinį tašką (μ_0, Σ_0):

$$\mu_0 = \frac{1}{K}\sum_{j=1}^{K}\alpha_j^0,$$

$$\Sigma_0 = \frac{1}{K}\sum_{j=1}^{K}\big(\alpha_j^0 - \mu_0\big)\big(\alpha_j^0 - \mu_0\big)^T,$$

čia $\alpha_j^0 = \ln\frac{P_j^0}{1-P_j^0}$, pasinaudojant paprastu santykinės rizikos įverčiu

$$P_j^0 = \frac{Y_j^0}{N_j}, \qquad P_j^0 \neq 0, \quad j = 1, \ldots, K.$$

4 Išvados

Darbe Bajeso metodas pritaikytas Puasono–Gauso kelių mažų dažnių tikimybių modelio parametrams įvertinti. Sudarytas algoritmas gali būti taikomas socialinių ir medicininių duomenų analizei.

Literatūra

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover Publications, New York, 1964.

[2] P.C. Bradley and A.L. Thomas. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall/CRC, 2000.

[3] D. Clayton and J. Kaldor. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, **43**(3):671–681, 1987.

[4] J.E. Dennis and R.B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. SIAM, 1996.

[5] L.V. Kantorovich and G.P. Akilov. *Functional Analysis*. Pergamon Press, New York, 1982.

[6] D.W. Pearce. *Aiškinamasis ekonomikos anglų–lietuvių kalbų žodynas*. TEV, Vilnius, 2006. Vertimas iš anglų k.

[7] R.K. Tsutakava, G.L. Shoop and C.J. Marienfield. Empirical Bayes estimation of cancer mortality rates. *Stat. Med.*, **4**(2):201–212, 1985.

SUMMARY

An algorithm for the assessment of several small rates

L. Sakalauskas, I. Vaiciulyte

The present paper describes the empirical Bayesian approach applied in the estimation of several small rates. Modeling by empirical Bayesian approach the probabilities of several rare events, it is assumed that the frequencies of events follow to Poisson's law with different parameters, which are correlated Gaussian random values. The unknown parameters are estimated by the maximum likelihood method computing the integrals appeared here by Hermite–Gauss quadratures. The equations derived that are satisfied by maximum likelihood estimates of model parameters.

Keywords: Bayesian approach, likelihood method, Poisson–Gaussian model.