Lietuvos matematikos rinkinys ISSN 0132-2818
Proc. of the Lithuanian Mathematical Society, Ser. A www.mii.lt/LMR/
Vol. **53**, 2012, 42–47

Empirical Bayesian regression model for estimation of small rates

Gintautas Jakimauskas, Leonidas Sakalauskas

Institute of Mathematics and Informatics, Vilnius University
Akademijos 4, LT-08663 Vilnius
E-mail: gintautas.jakimauskas@mii.vu.lt, leonidas.sakalauskas@mii.vu.lt

Abstract. The efficiency of adding an auxiliary regression variable to the logit model in estimation of small probabilities in large populations is considered. Let us consider two models of distribution of unknown probabilities: the probabilities have gamma distribution (model (A)), or logits of the probabilities have Gaussian distribution (model (B)). In modification of model (B) we will use additional regression variable for Gaussian mean (model (BR)). We have selected real data from Database of Indicators of Statistics Lithuania – Working-age persons recognized as disabled for the first time by administrative territory, year 2010 (number of populations $K = 60$). Additionally, we have used average annual population data by administrative territory. The auxiliary regression variable was based on data – Number of hospital discharges by administrative territory, year 2010. We obtained initial parameters using simple iterative procedures for models (A), (B) and (BR). At the second stage we performed various tests using Monte-Carlo simulation (using models (A), (B) and (BR)). The main goal was to select an appropriate model and to propose some recommendations for using gamma and logit (with or without auxiliary regression variable) models for Bayesian estimation. The results show that a Monte Carlo simulation method enables us to determine which estimation model is preferable.

Keywords: empirical Bayesian estimation, gamma model, logit model.

Introduction

Let us have K populations $A_1, A_2, \ldots, A_K$, consisting of N_j individuals, resp., and some event (e.g., some disease), can occur in these populations. We observe number of events $\{Y_j\} \stackrel{\text{def.}}{=} (Y_j,\ j = 1, 2, \ldots, K)$. We assume that number of events are caused by an unknown probabilities $\{\lambda_j\} \stackrel{\text{def.}}{=} (\lambda_j,\ j = 1, 2, \ldots, K)$, which are equal for each individual from the same population.

Let us consider the problem of estimation of small probabilities in large populations. The number of corresponding events depends on size of the population and the probability of the single event. It is assumed that number of events in each population has Poisson distribution with certain parameter.

It is assumed in empirical Bayesian estimation that the probabilities of events in populations are random and have some certain distribution. It is well known (see, e.g., [1, 4]) that Bayesian estimates of the unknown probabilities have substantially smaller mean square error as compared with mean square error of simple relative risk estimates $\bar{\lambda}_j^{RR} = Y_j/N_j,\ j = 1, 2, \ldots, K$. Note that if an assump-

tion that all probabilities are equal holds, we can use mean relative risk estimate $\{\bar{\lambda}_j^{MRR}\} \equiv \bar{\lambda}^{MRR} = \sum_{j=1}^K Y_j / \sum_{j=1}^K N_j$.

Let us consider three models of distribution of unknown probabilities: the probabilities have gamma distribution with shape parameter $\nu > 0$ and scale parameter $\alpha > 0$ (model (A)), logits of the probabilities have Gaussian distribution with mean μ and variance σ^2 (model (B)), modification of model (B) with auxiliary regression variable $\{Z_j\}$ (we use notation defined above), assuming that $\mu(j) = \mu_0 + \mu_1 Z_j$, $j = 1, 2, \ldots, K$, (model (BR)).

Comparison of model (A) and model (B) is given in [3]. In this paper we present some results of tests of model (BR) using Monte Carlo simulation based on real data from Database of Indicators of Statistics Lithuania (see http://www.stat.gov.lt/).

The main goal was to show the efficiency of using an auxiliary regression variable and to propose some recommendations for using gamma and logit models (with or without auxiliary regression variable) for Bayesian estimation.

1 Mathematical models

Let number of events $\{Y_j\}$ be a sample of independent random variables (r.v.'s) $\{\mathbf{Y}_j\} \stackrel{\text{def.}}{=} (\mathbf{Y}_j,\ j = 1, 2, \ldots, K)$, with binomial distribution (resp., with parameters $(\lambda_j,\ N_j)$, $j = 1, 2, \ldots, K$. Clearly, $\mathbf{E}\mathbf{Y}_j = \lambda_j N_j$, $j = 1, 2, \ldots, K$.

An assumption is often made (see, e.g., [1, 5]) that r.v.'s $\{\mathbf{Y}_j\}$ have a Poisson distribution with parameters $\lambda_j N_j$, $j = 1, 2, \ldots, K$, i.e.

$$\mathbf{P}\{\mathbf{Y}_j = m\} = h(m, \lambda_j N_j), \quad m = 0, 1, \ldots,\ j = 1, 2, \ldots, K,$$
$$h(m,\ z) \stackrel{\text{def.}}{=} e^{-z}\frac{z^m}{m!}, \quad m = 0, 1, \ldots,\ z > 0.$$

We will consider the mathematical model assuming that unknown probabilities $\{\lambda_j\}$ are independent identically distributed (i.i.d.) r.v.'s with distribution function F from the certain class $\mathcal{F}$. Our problem is to get estimates of unknown probabilities $\{\hat{\lambda}_j\}$ from the observed number of events $\{Y_j\}$, assuming that $F \in \mathcal{F}$.

Model (A). Let us make an assumption that $\{\lambda_j\}$ are i.i.d. gamma r.v.'s with shape parameter $\nu > 0$ and scale parameter $\alpha > 0$, i.e. the distribution function F has the distribution density

$$f(x) = f(x; \nu, \alpha) = \frac{\alpha \cdot (\alpha \cdot x)^{\nu-1}}{\Gamma(\nu)} e^{-\alpha x}, \quad 0 \leqslant x < \infty.$$

Then $\mathbf{E}\lambda_j = \nu/\alpha$, and $\mathbf{D}\lambda_j = \nu/\alpha^2$. Moreover,

$$\mathbf{E}(\lambda_j \mid \mathbf{Y}_j = Y_j) = \frac{Y_j + \nu}{N_j + \alpha}, \quad j = 1, 2, \ldots, K. \tag{1}$$

Empirical Bayes estimate $\{\hat{\lambda}_j\}$, which is certain compromise between mean relative risk estimate $\{\bar{\lambda}_j^{MRR}\}$ and relative risk estimate $\{\bar{\lambda}_j^{RR}\}$ is obtained by (1) using parameter estimates $(\hat{\nu}, \hat{\alpha})$.

Model (B). Alternatively, we will consider Bayes estimate $\tilde{\lambda}_j$, which is obtained under assumption that unknown probabilities are i.i.d. r.v.'s such that their logits

$\alpha_j = \ln(\lambda_j/(1-\lambda_j))$, $j = 1, 2, \ldots, K$, are i.i.d. Gaussian r.v.'s with mean μ and variance σ^2 and corresponding distribution density φ.

Model (BR). Additionally, let us introduce an auxiliary regression variable $\{Z_j\}$, assuming that $\mu(j) = \mu_0 + \mu_1 Z_j$, $j = 1, 2, \ldots, K$. This variable is considered non-random, so all formulae for model (B) hold also for model (BR).

In the case of model (B) and model (BR) conditional expectation of $\{\lambda_j\}$ has the following form (see, e.g., [2]):

$$\mathbf{E}(\lambda_j \mid \mathbf{Y}_j = Y_j) = \frac{\int_{-\infty}^{\infty} \frac{1}{1+e^{-x}} h\big(Y_j, \frac{N_j}{1+e^{-x}}\big)\varphi(x;\ \mu, \sigma^2)\, dx}{D_j(\mu, \sigma^2)}, \quad j = 1, 2, \ldots, K, \tag{2}$$

$$D_j(\mu, \sigma^2) = \int_{-\infty}^{\infty} h\left(Y_j, \frac{N_j}{1+e^{-x}}\right)\varphi(x; \mu, \sigma^2)\, dx, \quad j = 1, 2, \ldots, K.$$

For model (A) maximum likelihood function has the following form:

$$L_A(\nu, \alpha) = \sum_{j=1}^{K} \left(\ln \frac{\Gamma(Y_j + \nu)}{\Gamma(\nu)} + \nu \ln(\alpha) - (Y_j + \nu)\ln(N_j + \alpha) + Y_j \ln N_j \right). \tag{3}$$

For models (B) and (BR) maximum likelihood function has the following form:

$$L_B(\mu, \sigma^2) = \sum_{j=1}^{K} \ln D_j\big(\mu, \sigma^2\big). \tag{4}$$

Maximum likelihood estimates $\{\hat{\lambda}_j\}$ and $\{\tilde{\lambda}_j\}$ are obtained by maximizing (3), resp., (4) and replacing parameter values in (1) or (2) with $(\hat{\nu}, \hat{\alpha})$ or $(\tilde{\mu}, \tilde{\sigma}^2)$.

2 Computer simulation results

As initial data for the simulation $\{Y_j\}$ we have selected real data from Database of Indicators of Statistics Lithuania (Table M3140706) – Working-age persons recognized as disabled for the first time by administrative territory, year 2010 (number

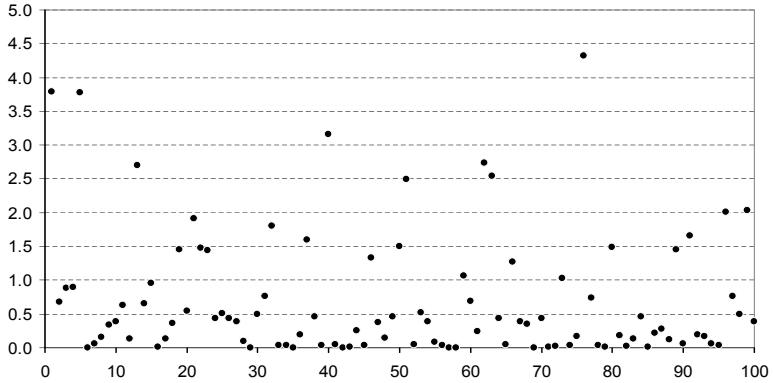

Fig. 1. $L_{BR} - L_B$ (dataset 6, simulation by model (A)).

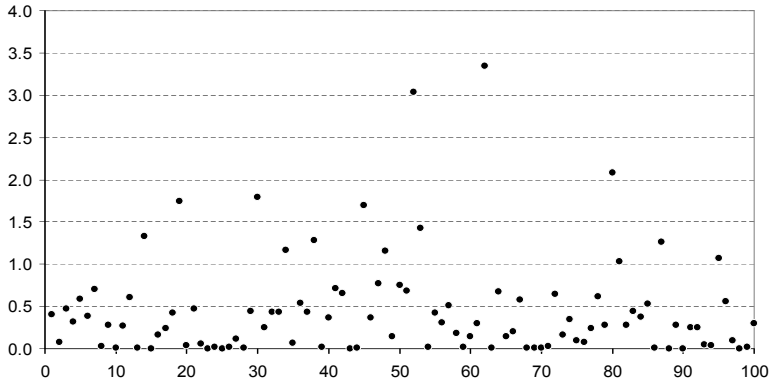

Fig. 2. $L_{BR} - L_B$ (dataset 6, simulation by model (B)).

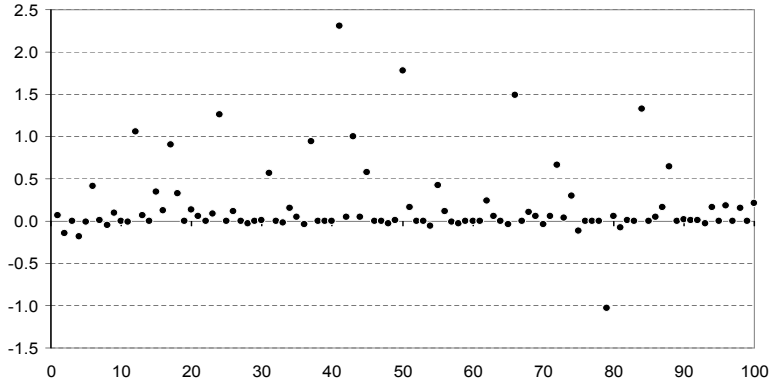

Fig. 3. $L_A - L_B$ (dataset 6, simulation by model (B)).

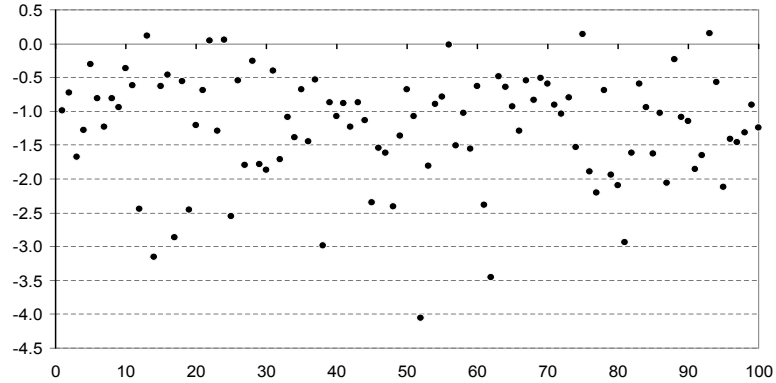

Fig. 4. $L_A - L_{BR}$ (dataset 6, simulation by model (B)).

of populations ($K = 60$), 9 data sets in total (total 15432 cases). We have used corresponding average annual population data (Table M3010211, total population 3286820) by administrative territory N_j, $j = 1, 2, \ldots, K$.

The auxiliary regression variable was based on data – Number of hospital discharges by administrative territory (Table M3140312), year 2010. Values of Z_j, $j = 1, 2, \ldots, K$, were obtained by dividing number of hospital discharges by population data in each administrative territory.

At the initial stage real sample data were evaluated using model (A) and model (B), i.e. we obtained starting estimates $\{\hat{\lambda}_j\}_0$ and $\{\tilde{\lambda}_j\}_0$. For model (BR) we started with parameter $m_1 = 0$. For details see [3].

Having the starting estimate, we simulated random realizations using Monte Carlo method (usually 100 independent realizations), using model (A) and model (B). In each model case, for simulated realizations we calculated an estimate $\{\hat{\lambda}_j\}$, estimate $\{\tilde{\lambda}_j\}$, and values of ML function $L_A(\hat{\nu}, \hat{\alpha})$, $L_B(\tilde{\mu}, \tilde{\sigma}^2)$ and $L_{BR}(\tilde{\mu}, \tilde{\sigma}^2)$. For details on optimization algorithm see [3].

The main goal was to compare efficiency of model (BR) with efficiency of model (B) by difference $L_{BR} - L_B$. The results show that for some data sets this difference is big enough to prefer model (BR) to model (A), in cases when model (A) is preferable to model (B). For some simulation results see Figs. 1–4.

References

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

[2] R. Gurevičius, G. Jakimauskas and L. Sakalauskas. Empirical Bayesian estimation of small mortality rates. In M. Grasserbauer, L. Sakalauskas and E.K. Zavadskas(Eds.), *Proc. of the 5th International Vilnius Conference [and] EURO-mini Conference „Knowledge-based Technologies and OR Methodologies for Decisions of Sustainable Development* (*KORSD-2009*)*, Vilnius, Lithuania, 2009*, pp. 290–295, Vilnius, 2009. Technika.

[3] G. Jakimauskas. Gamma and logit models in empirical Bayesian estimation of probabilities or rare events. In L. Sakalauskas, A. Tomasgard and S.W. Wallace(Eds.), *Proc. of the International Workshop StoProg-2102* (*Stochastic Programming for Implementation and Advanced Applications*)*, Neringa, Lithuania, 2012*, pp. 43–48, Vilnius, 2012. The Association of Lithuanian Serials.

[4] J.L. Meza. Empirical Bayes estimation smoothing of relative risks in disease mapping. *J. Stat. Plann. Inf.*, **112**:43–62, 2003.

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

REZIUMĖ

Empirinis Bayeso regresinis modelis mažų tikimybių vertinimui

G. Jakimauskas, L. Sakalauskas

Nagrinėjamas papildomo regresijos kintamojo pridėjimo prie logit modelio efektyvumas vertinant mažas tikimybes didelėse populiacijose. Nagrinėjami du nežinomų tikimybių pasiskirstymo modeliai: tikimybės pasiskirsčiusios pagal gama skirstinį (modelis (A)), arba tikimybių logit'ai turi Gauso skirstinį (modelis (B)). Modifikuotame modelyje (B) naudojamas papildomas regresijos kintamasis Gauso skirstinio vidurkiui (modelis (BR)). Naudojami realūs duomenys iš Lietuvos Statistikos departamento Duomenų bazės – Darbingo amžiaus asmenys, pirmą kartą pripažinti neįgaliaisiais pagal administracinę teritoriją, 2010 metai (teritorijų skaičius $K = 60$). Be to, naudojami vidutinio metinio gyventojų skaičiaus duomenys pagal administracinę teritoriją. Papildomas regresijos kintamasis paremtas duomenimis – Ligoninėse gydytų ligonių skaičius, 2010 metai. Buvo gauti pradiniai

parametrai naudojant paprastas iteracines procedūras modeliams (A) ir (B). Antrame etape buvo atlikti įvairūs testai naudojant Monte Carlo modeliavimą (naudojant modelius (A), (B) ir (BR). Pagrindinis tikslas buvo parinkti tinkamiausią modelį ir duoti rekomendacijas naudoti gama ar logit modelį (su ar be papildomu regresijos kintamuoju) Bayeso vertinimui. Rezultatai rodo, kad Monte-Carlo modeliavimas įgalina parinkti tinkamesnį vertinimo modelį.

Raktiniai žodžiai: empirinis Bayeso vertinimas, gama modelis, logit modelis.