ANALYSIS OF MEDICAL DATA BY EMPIRICAL BAYES METHOD

L. Sakalauskas, G. Jakimauskas, J. Susinskas
Institute of Mathematics and Informatics
Vilnius, LITHUANIA
e-mail: sakal@ktl.mii.lt

Abstract

The empirical Bayesian analysis of rare rates is considered in the paper. The condition for non-singularity of Bayesian estimates is given and the clustering algorithm is developed, using the property of the Poisson–Gaussian model to treat probabilities of events in populations being the same, if the variance of probabilities is small. The approach developed is applied to the analysis of dissimilarities of homicide and suicide data in Lithuania, 2003–2004.

1 Introduction

The Bayes method is widespread in the reliable analysis of spatial information, because it can evaluate rates of certain events not only from the current population data, but also from data of other populations (Tsutakava et al. (1985), Quigley et al. (2007), etc.). Empirical Bayesian estimates are shown to have substantially smaller mean squared errors than RR estimates. Typical observables in risk mapping are numbers of events that obey Poisson distribution, depending on the event rate and the observation time for each population. In this paper, numerical features of empirical Bayesian estimation techniques for the Poisson–Gaussian model are addressed, when prior distribution of logits is normal with the parameters estimated by the maximal likelihood (ML) method (Tsutakava et al. (1985), Sakalauskas (2009)). The nonsingularity conditions are derived in estimating the parameters of prior distribution. Thus, since the emprical Bayes approach for the Poisson–Gaussian model distinguishes by the property to treat probabilities of events in populations being the same, when the numbers of events are not varying much, the clustering .algorithm is developed based on this property. We utilize a Lithuanian mortality data set of 2003–2004 to estimate the underlying true risks and show the applicability of the approach considered.

2 Poisson–Gaussian Model

Let us consider a set $\Lambda = (A_1, A_2, \ldots, A_K)$ of K populations, where each population A_j consists of N_j individuals. Assume that some event (e.g., death due to some disease) can occur in the populations under observation. The aim is to estimate unknown probabilities of eventsP_j, when the numbers Y_j of events in populations are observed, $j = \overline{1, K}$. Since a simple estimate of the relative risk $\frac{Y_j}{N_j}$ not useful in many cases due to great differences in population size N_j, the empirical Bayesian approach is applied. An

assumption is often corroborated (Tsutakava (1985), etc.) that the numbers of cases Y_j follows the Poisson distribution with the parameters$\lambda_j = N_j \cdot P_j$, i.e.:

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

It is of interest to consider the model, in which the logits

$$\alpha_j = \ln \frac{P_j}{1 - P_j} \tag{2}$$

are normally distributed with the parametersμ, σ. Thus, the density of logit (2) is

$$g(\alpha_j, \mu, \sigma) = \frac{\exp\big(-\frac{(\alpha_j - \mu)^2}{2\sigma^2}\big)}{\sqrt{2\pi}\sigma}. \tag{3}$$

Then the rates P_j are evaluated as a posteriori means for given μ, σ,

$$P_j = \frac{\int_{-\infty}^{\infty} \frac{1}{1+e^{-\alpha}} f\big(Y_j, \frac{N_j}{1+e^{-\alpha}}\big) g(\alpha, \mu, \sigma) d\alpha}{D_j(\mu, \sigma)}, \tag{4}$$

where

$$D_j(\mu, \sigma) = \int_{-\infty}^{\infty} f\left(Y_j, \frac{N_j}{1 + e^{-\alpha}}\right) g(\alpha, \mu, \sigma) d\alpha \tag{5}$$

is the a posteriori probability of the number of event in the jth population, $j = \overline{1, K}$.

In the emprical Bayessian approach the unknown parameters μ, σ are estimated by the maximal likelihood method (Tsutakava et al. (1985)). The logarithmic likelihood function after some manipulations, is as follows:

$$L(\mu, \sigma) = -\sum_{j=1}^{K} \ln \left(\int_{-\infty}^{\infty} f\left(Y_j, \frac{N_j}{1 + e^{-\alpha}}\right) g(\alpha, \mu, \sigma) d\alpha \right) = -\sum_{j=1}^{K} \ln \big(D_j(\mu, \sigma)\big), \tag{6}$$

which has to be minimized to get estimates for the parametersμ, σ. The likelihood function (6) is differentiable many times with respect to the parametersμ, σ. Equating the derivatives of the logarithmic likelihood function to zero the equations are derived, which the ML estimates of μ and σ should satisfy (see details in Sakalauskas (2009)):

$$\mu = \frac{1}{K} \sum_{j=1}^{K} \frac{\int_{-\infty}^{\infty} \alpha f\big(Y_j, \frac{N_j}{1+e^{-\alpha}} big\big) g(\alpha, \mu, \sigma) d\alpha}{D_j(\mu, \sigma)}, \tag{7}$$

$$\sigma^2 = \frac{1}{K} \sum_{j=1}^{K} \frac{\int_{-\infty}^{\infty} (\alpha - \mu)^2 f\big(Y_j, \frac{N_j}{1+e^{-\alpha}}\big) g(\alpha, \mu, \sigma) d\alpha}{D_j(\mu, \sigma)}. \tag{8}$$

However, solution of these equations exists only under the nonsingularity assumption of the ML estimate of σ (i.e., $\sigma^2 > 0$). The solution of equations (7), (8) exists if

$$\sum_{j=1}^{K} (Y_j - N_j \cdot P)^2 > \sum_{j=1}^{K} Y_j. \tag{9}$$

Otherwise, the ML estimates are

$$\mu = \ln \frac{P}{1-P}, \quad \sigma = 0, \tag{10}$$

where

$$P = \sum_{j=1}^{K} Y_j \Big/ \sum_{j=1}^{K} N_j. \tag{11}$$

It follows from condition (9) that the singularity occurs most often in small populations. Hence, this condition may be used to establish a population set with rare events. It is easy to make sure that in the case of singularity (i.e., $\sigma = 0$) the numbers of event remain constant for all the populations, that is $P_j \equiv P$. The corresponding value of the ML function is

$$L(\mu*, 0) = \sum_{j=1}^{K} \big(N_j \cdot P - Y_j \cdot \ln(N_j \cdot P)\big) = \sum_{j=1}^{K} Y_j \cdot \big(1 - \ln(N_j \cdot P)\big). \tag{12}$$

3 Application in Clustering

The property derived of the Poisson–Gaussian model to treat populations with relative ratios, which are close each to other as having the same probabilities of events may be applied to map clustering, too. Let us consider a set of clusters Ξ consisting of the populations of the set $\Lambda = (A_1, A_2, \ldots, A_K)$. Note, that we treat the subsets of contiguous populations as clusters (i.e., any population in a cluster has a common border with some other population from this cluster), in which the condition of zero variance derived from (9) is true:

$$\sum_{A_j \in C_\delta} (Y_j - N_j \cdot P)^2 - Y_j \le 0. \tag{13}$$

Let $C = (C_1, C_2, \ldots, C_M)$ be a set clusters that covers the whole set of populations $\Lambda : \cap_{i=1}^{M} C_i = \Lambda$, $C_i \cap C_j = \emptyset$, $i \ne j$, $i, j = \overline{1, M}$. We select the clustering set so that the likelihood function (6) becomes minimal: Thus, using (12) after some simple manipulations one may make sure that the best clustering set should provide the minimum of the function

$$\Pi(C) = \sum_{j=1}^{M} \sum_{A_\delta \in C_j} Y_\delta \cdot \ln \left(\frac{\sum_{A_\delta \in C_j} Y_\delta}{\sum_{A_\delta \in C_j} N_\delta} \right) \to \min_{C}. \tag{14}$$

The corresponding probabilities of the events for the populations of the cluster are the same:

$$P_j = \frac{\sum_{\delta \in C_j} Y_\delta}{\sum_{\delta \in C_j} N_\delta}. \tag{15}$$

Note, that the number of possible clusters is rather large and we have to look through huge number of clusters, when the clustering set should be established with respect to (14). However heuristic simplifications may be applied using the next proposition.

Proposition 1 *Let C_1 and C_2 be two populations with numbers of events Y_1, Y_2 and sizes N_1, N_2. Then*

$$Y_1 \cdot \ln\left(\frac{Y_1+Y_2}{N_1+N_2}\cdot N_1\right) + Y_2 \cdot \ln\left(\frac{Y_1+Y_2}{N_1+N_2}\cdot N_2\right) \le Y_1 \cdot \ln(Y_1) + Y_2 \cdot \ln(Y_2). \quad (16)$$

The proof of the proposition is simple and performed by elementary manipulations.

Thus, it follows from (16) that merging of two clusters causes the ML function to decrease. This property can be used for a simplified search of the best clustering set. We start from the initial clustering set, consisting of K clusters, each having only one population. The next two clusters are merged, if condition (13) remains valid in the merged cluster and the decrease of ML function is minimal among all the possible merging combinations, and this procedure is repeated until termination.

4 Implementation and Discussion

The method developed was applied to analysis of data on homicide and suicide mortality in Lithuania in 2003/2004 (all the events in population, for men and women). Integration and minimization of the likelihood function was performed by means of the mathematical software MATHCAD. We can see the decrease in variance of empirical Bayesian estimates with a comparison to RR. The empirical Bayesian estimation enables us to observe certain spatial effects in the distribution of suicide rates in populations. The singularity of empirical Bayesian analysis with the Poisson–Gaussian model often occurs while analyzing real data. In this paper, we derive a condition of non-singularity for the empirical Bayesian method. The property of the empirical Bayesian approach to treat populations with ratios, which are close each to other, is discussed through an application of populations clustering. The approach developed has been applied in the analysis of social and medical data, and its simplicity and applicability is approved.

References

[1] Sakalauskas L. (1995). On Bayes analysis of small rates in medicine. In: *Proc. of the Intern. Conf. "Computer Data Analysis and Modeling"*, September 14–19, 1995, Minsk. Vol. **1**, pp. 127–130.

[2] Tsutakava R.K., Shoop G.L., Marienfield C.J. (1985). Empirical Bayes estimation of cancer mortality rates. *Statistics in Medicine*. Vol. **4**, pp. 201–212.