IMPLEMENTATION OF THE POISSON-GAUSSIAN REGRESSION MODEL IN EMPIRICAL BAYES ESTIMATION OF SMALL PROBABILITIES

G. Jakimauskas[1], L. Sakalauskas[2]
Institute of Mathematics and Informatics, Vilnius University
Vilnius, LITHUANIA
e-mail: [1]gintautas.jakimauskas@mii.vu.lt, [2]leonidas.sakalauskas@mii.vu.lt

Abstract

The problem of implementation of the Poisson-Gaussian regression models in empirical Bayesian estimation of the small probabilities is considered. A bootstrap method using Monte-Carlo simulation is proposed. The method is applied to real-world USA cancer data combined with some possible regression variables, assuming they may have influence on the actual cancer data.

1 Introduction

Let us consider the problem of probability estimation of rare events in large populations (e.g., probabilities of some disease, homicides, suicides, etc.). The respective number of events depends on the population size and on the probability of a single event. Let us assume that probability of a single event depends only on population and these probabilities are the same in the same population. Moreover, assume that all events in all populations are independent. Under such assumptions number of events in each population will follow the Bernoulli distribution.

An event count refers to the number of times an event occurred in specific population. The benchmark model for count data is the Poisson distribution.

The Poisson distribution is the simplest distribution for modeling count data. However, it has one obvious limitation: its variance is equal to its mean. In case of real data we usually have so-called overdispersion: empirical variance is significantly bigger than empirical mean. In this case we can add some independent mixing distribution which increases variance of the combined distribution. By selecting parameters of the mixing distribution we can adjust the mean and the variance of the combined distribution to the empirical mean and the empirical variance of the real data. The simplest model adds gamma distribution to the Poisson distribution. The resulting distribution is known as negative binomial distribution or Poisson-gamma distribution. This distribution is more dispersed than the Poisson distribution. Obviously, negative binomial distribution can accommodate overdispersion but not underdispersion. There are many generalizations of the Poisson distributions (see, e.g., [2], [3], [6]).

Count data regression models have a widespread use (see, e.g., [2], [6]). The mean parameter of the Poisson-gamma model is usually parametrized using exponential link function of the regressors, in order to ensure that mean parameter is strictly greater than zero. As an alternative to the Poisson-gamma distribution, we will consider

Poisson-Gaussian distribution (see, e.g., [7], [8]). In this case the additional link function is not needed, and adding regression variables is very simple and clear. However, the calculations using Poisson-Gaussian model are much more complicated, and we need to use Hermite-Gauss numerical integration formulae (see, e.g. [1]).

2 Mathematical models

Let observed number of events $\{Y_j\} = Y_j,\ j = 1, \ldots, K$, be a sample of independent random variables $\{\mathbf{Y}_j\}$ with binomial distribution, respectively, with number of experiments $\{N_j\}$ and success probabilities $\{\lambda_j\}$. Clearly, $\{\mathbf{E}(\mathbf{Y}_j)\} = \{\lambda_j N_j\}$.

An assumption is often made (see, e.g., [5], [8]) that random variables $\{\mathbf{Y}_j\}$ have a Poisson distribution with parameters, respectively, $\{\lambda_j N_j\}$, i.e.

$$\mathbf{P}\{\mathbf{Y}_j = m\} = h(m,\ \lambda_j N_j),\ m = 0, 1, \ldots,\ j = 1, \ldots, K,$$

where

$$h(m,\ z) = e^{-z}\frac{z^m}{m!},\ m = 0, 1, \ldots,\ z > 0.$$

We will consider the mathematical model assuming that unknown probabilities $\{\lambda_j\}$ are independent identically distributed random variables with distribution function F from the certain class of distribution functions $\mathcal{F}$. Our problem is to get empirical Bayes estimates (see, e.g., [4]) of unknown probabilities $\{\hat{\lambda}_j\}$ from the observed number of events $\{Y_j\}$, assuming that $F \in \mathcal{F}$.

Poisson-gamma model. Given population sizes $\{N_j\}$, let random variables $\{\mathbf{Y}_j\}$ have a Poisson distribution with parameters, respectively, $\{\lambda_j N_j\}$, where $\{\lambda_j\}$ are independent identically distributed gamma random variables with shape parameter $\nu > 0$ and scale parameter $\alpha > 0$, i.e. the distribution function F has the distribution density

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

Then $\mathbf{E}(\lambda_j) = \nu/\alpha$, and $\mathbf{E}(\lambda_j - \mathbf{E}(\lambda_j))^2 = \nu/\alpha^2,\ j = 1, \ldots, K$. Given observed number of events $\{Y_j\}$ and population sizes $\{N_j\}$, Bayes estimates for $\{\lambda_j\}$ are (see, e.g. [5])

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

Corresponding maximum likelihood function for parameters (ν, α) is

$$L(\nu, \alpha) = \sum_{j=1}^{K} \Big(\ln \frac{\Gamma(Y_j + \nu)}{\Gamma(\nu)} + \nu \ln(\alpha) - \\ -(Y_j + \nu)\ln(N_j + \alpha) + Y_j \ln N_j \Big). \tag{2}$$

Empirical Bayes estimates $\{\hat{\lambda}_j\}$ are obtained by maximizing (2) and replacing parameters (ν, α) in (1) with obtained parameters $(\hat{\nu},\ \hat{\alpha})$.

Poisson-Gaussian model. 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_{μ,σ^2}.

Poisson-Gaussian model with regression variables. Additionally, let us introduce an auxiliary regression variables $\{Z_j\}^{(s)}$, $s = 1, \ldots, M$, assuming that $\mu(j) = \mu_0 + \mu_1 Z_j^{(1)} + \mu_2 Z_j^{(2)} + \ \cdots + \mu_M Z_j^{(M)}$, $j = 1,\ 2,\ \ldots\ , K$ (for our purposes we consider only simplified model without interactions of the regression variables). These variables are considered non-random, so all formulae for Poisson-Gaussian model hold also for Poisson-Gaussian model with regression variables.

In the case of both Poisson-Gaussian models conditional expectation of $\{\lambda_j\}$ has the following form:

$$\mathbf{E}(\lambda_j \mid \mathbf{Y}_j = Y_j) = D_j^{-1}(\mu(j), \sigma^2) \int\limits_{-\infty}^{\infty} \frac{1}{1+e^{-x}} h\left(Y_j, \frac{N_j}{1+e^{-x}}\right) \varphi_{\mu(j),\sigma^2}(x)\ dx,$$
$$j = 1, \ldots, K,$$

where

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

3 Implementation of the Poisson-Gaussian regression model

To demonstrate the implementation of the Poisson-Gaussian regression model the main intention of data selection was to select freely available datasets, preferably of certain relatively large population, from the trusted databases. We have selected real data from the database of the USA National Cancer Institute, years 2011 and 2012, number of administrative territories (states) $K = 50$, 23 datasets in total. Also we have used population data by administrative territories from the United States Census Bureau.

As a basis for the regression variables we have used corresponding real data by administrative territories (states) from the Health Indicators Warehouse of the USA Center for Disease Control and Prevention. We have analysed three of possible regression variables, assuming they may have influence on the actual cancer data: (1) "Depression Medicare beneficiaries", (2) "High cholesterol Medicare beneficiaries", (3) "Toxic chemicals (pounds)".

For each dataset we have performed Monte-Carlo computer simulation of (typically) 100 independent realizations using both Poisson-gamma and Poisson-Gaussian models with corresponding parameters estimated from the real data (assuming that there are no regression). At the next stage we estimated (using maximum likelihood

method) Poisson-Gaussian model parameters without regression variables, and, alternatively, Poisson-Gaussian model parameters with selected regression variables. In the process, corresponding values of the maximum likelihood function were obtained. This procedure was applied to the real data, and to the 100 simulated realizations (either Poisson-gamma model realizations or Poisson-Gaussian model realizations).

The key point of this method is comparing difference of values of maximum likelihood function (for model with regression variables and for model without regression variables) for the real data with analogous differences for simulated realizations. Because simulated realizations have no influence of regression variables, they only have small random differences of values of maximum likelihood function, which main characteristics can be easily calculated. As a simple method, we can apply 3σ rule to detect presence of the regression variables.

As expected, the simulation results did not show significant difference of simulation using Poisson-gamma and Poisson-Gaussian models. Implementing the simple 3σ rule, we have found that for datasets 4, 9, 10, 16, 18 it is recommended to use regression variable "High cholesterol Medicare beneficiaries" for empirical Bayes estimation. For datasets 2, 9, 10, 16 it is recommended to use regression variable "Depression Medicare beneficiaries" for empirical Bayes estimation. As for regression variable "Toxic chemicals (pounds)" (combined with population size or with area size), we did not find influence of this variable.

References

[1] Abramovich M., Stegun I.A. (1968). *Handbook of Mathematical Functions*. Dover, New York.

[2] Cameron A.C., Trivedi P.K. (1998). *Regression Analysis of Count Data*. University Press, Cambridge.

[3] Cameron A.C., Trivedi P.K. (2005). *Microeconometrics. Methods and Applications*. University Press, Cambridge.

[4] Carlin B.P., Louis T.A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall, London.

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

[6] Hilbe J.M. (2011). *Negative Binomial Regression*. University Press, Cambridge.

[7] Sakalauskas L. (2010). On the Empirical Bayesian Approach for the Poisson-Gaussian Model. *Methodology and Computing in Applied Probability*. Vol. **12**, Issue **2**, pp. 247-259.

[8] 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.