Statistika, Vol. 14 No. 1, 1 – 4
Mei 2014

Koreksi Penduga SMR dalam Disease Mapping

SEPTIADI PADMADISASTRA, JADI SUPRIJADI

Jurusan Statistika Unpad Bandung
s_padmadisastra@yahoo.com

ABSTRAK

Dalam makalah ini dibahas penyusunan peta penyakit (Disease mapping) dengan memperhitungkan adanya kasus data yang tidak tercatat (underreported). Data yang diamati adalah banyaknya penderita suatu penyakit disejumlah wilayah kecil disebuah kotaseperti kecamatan-kecamatan. Kasus ini menyebabkan maximum likelihood estimator untuk parameter resiko relative (SMR) tidak dapat dicari. Oleh karenanya dalam makalah ini diusulkan sebuah metode Bayesian dengan data underreported.
Kata Kunci: maximum likelihood estimator, disease mapping, Bayesian, penduga SMR

1. PENDAHULUAN

Peta penyakit (*disease mapping*) merupakan sebuah visualisasi dari distribusi resiko relative (θ) suatu penyakit dalam sebuah wilayah. Peta sangat bermanfaat bagi pemerintah dalam mengalokasikan anggaran kesehatan dalam rangka menanggulangi penyebaran penyakit, selain itu juga berguna dalma membentuk sebuah aetyologi hipotesis mengenai sebuah penyakit. Oleh karenanya, agar keputusan yang dibuat berdasarkan peta ini dapat dipertanggung jawabkan, maka data yang menjadi dasar penyusunannya haruslah akurat; mencerminkan keadaan sebenarnya.

Data yang dijadikan dasar penyusunannya merupakan sebuah data hasil pencacahan; yaitu mengenai jumlah penderita. Sebagaimana sudah banyak dilaporkan, misalnya Scollinik (2002), bahwa data hasil cacahan sering mengalami gangguan (rusak), disebut sebagai *damaged* atau *underreportedvariable*. Artinya nilai yang dilaporkan tidak sesuai dengan seharusnya. Beragam penyebab tentang kejadian ini,terkait dengan kasus yang diteliti sekarang adalah kasus tidak lengkapnya catatan mengenai jumlah penderita penyakit DBD di kota Bandung. Penyebab tidak kengkapnya catatan adalah semakin banyaknya tempat pengobatan alternative yang melayani penyembuhan beragam penyakit. Jumlah mereka yang berkunjung tentunya tidak tercatat di dinas kesehatan kota. Akibatnya data yang ada di dinas jumlahnya tidak sesuai dengan kenyataannya, lebih rendah dari yang seharusnya; rusak atau *underreported*.

Kejadian ini tentu perlu mendapat perhatian dalam menyusun peta penyebaran penyakit, teristimewa dalam menetapkan penaksir resiko relative($\hat{\theta}$), disebuah wilayah yang akan dipetakan. Dalam disease mapping, diassumsikan bahwa banyak kasus teramati (O_i) berdistribusi Poisson dengan parameter $E\theta$,E merupakan banyak kasus yang diharapkan, jadi disebuah wilayah I berlaku:

$$P(O_i = o) = \frac{(E_i\theta_i)^o \exp(-E_i\theta_i)}{o!}, \quad o = 0,1,\ldots \tag{1}$$

Penaksir ini *maximum likelihood estimator* (MLE) untuk parameter θ_i,yaitu penaksir resiko relative, merupakan sebuah ratio antara banyak kasus teramati (O_i) dengan banyak kasus yang diharapkan (E_i) :

$$\widehat{\theta_i} = \frac{O_i}{E_i}$$

dan bersifat tak bias, Lawson (2000). Penaksir ini disebut sebagai *Standardized Mortality Ratio* (SMR). Akan tetapi sebagai mana diketahui ada beberapa kekurangan dari penaksir ini, diantaranya dipengaruhi jumlah penduduk dalam wilayah;dalam hal wilayah dengan jumlah penduduk berbeda tetapi memiliki jumlah penderita yang sama. Wilayah dengan jumlah

penduduk besar akan memiliki nilai $\hat{\theta}$kecil dibandingkan dengan wilayah berpenduduk sedikit. Selain itu varians penaksir juga terpengaruh dengan akibat yang sama.

Oleh karena kelemahan ini maka orang mengusulkan sejumlah alternative metode, diantaranya *smoothing*, *Linear Bayes*, *Bayesian*, dan *Empirical Bayes*. Sebuah study yang dilakukan Lawson(2000) mengenai kesesuain metode-metode ini dalam medeskripsikan data hasil simulasi yang melibatkan beragam model resiko relative. Salah satu kesimpulannya adalah bahwa metode Bayesian (gamma – Poisson) adalah yang paling *robust* dalam semua model resiko relative yang dipelajari.

2. MODEL UNDERREPORTED

Pembahasan mengenai data dalam keadaan *Underreported* atau *damaged* sudah dikemukakan oleh Rao dan Rubin (1964) dalam konteks karakterisasi distribusi Poisson, dan Rao (1965) yang membahas inferensi parameter distribusi Poisson apabila ada kerusakan dalam data pengamatan.

Data dikatakan rusak atau *underreported* apabila nilai data sesungguhnya, n, hanya dilaporkan sebagain saja yaitu yang tercatatnya saja (o). Jadi ada bagian dari n yang tidak tercatat (u). Oleh karenanya $n = o + u$.Dan distribusi peluang bersyarat $P(O = o|N = n)$ disebut sebagai distribusi *survival*. Sebuah postulat dalam Rao dan Rubin, dikenal sebagai Rao-Rubin *condition*, mengatakan bahwa bila distribusi *survival* binomial dengan parameter n dan p (peluang tercatat) maka distribusi peluang n haruslah Poisson,karena

$$P(O = o) = P(O = o|N = O) = P(O = o|N > O)$$

Jika dan hanya jika n berdistribusi Poisson, dan marjinal dari o, merupakan bagian tercatat dari n, juga Poisson.

Proses inferensi mengenai parameter-parameter dalam distribusi Poisson dengan kasus underreported disampaikan oleh Charnet (2004). Sedangkan Scollnik menerapkannya dalam bidang assuransi dengan anggapan bahwa pengamatan bearsal dari distribusi Poisson dan kerusakan mengikuti pola binomial.

Dalam *disease mapping*, apabila n berdistribusi Poisson dengan parameter$\lambda = E\theta$,maka untuk sebuah pengamatan di wilayah i

$$P(N_i = n|E_i, \theta_i) = \frac{(E_i\theta_i)^n \exp(-E_i\theta_i)}{n!}, \quad n = 0,1, \ldots \tag{2}$$

dan apabIla peluang bahwa sesorang tercatat adalah p maka tentunya o bersyarat kepada n adalah binomial dengan parameter n dan p

$$P(O_i = o|n) = \binom{n}{o} p^o (1-p)^{n-o} \tag{3}$$

dan marjinal dari distribusi untuk o adalah Poisson dengam parameter ($pE\theta$), maka di wilayah i

$$P(O_i = o) = \sum_{n \geq o}^{\infty} P(O_i = o|n) P(N_i = n)$$

Diselesaikan menghasilkan:

$$P(O_i = o) = \frac{(p_i E_i \theta_i)^o \exp(-p_i E_i \theta_i)}{o!}, \quad o = 0,1, \ldots \tag{4}$$

Kalau distribusi (4) dipakai sebagai bentuk data pengamatan seperti dalam kasus data lengkap (sebutlah standar), penaksiran parameter p dan θ dalam model (4) memakai metode *maximum likelihood* menghasilkan persamaan normal singular, karenanya tidak ada jawab. Masalah lainnya,seperti dalam kasus data lengkap bahwa untuk menghindari kelemahan dari penaksir SMR, maka resiko relative tidak dipandang sebagai sebuah konstanta tetapi sebagai variable acak. dengan distribusi peluang tertentu. Jadi digunakan pendekatan Bayesian untuk memperbaikinya.

3. BAYESIAN UNDERREPORTED

Seperti diperoleh Lawson (2000) melalui simulasi, bahwa model dengan prior gamma (a, β) untuk pemetaan ternyata merupakan sebuah model yang *robust*. Berdasarkan pertimbangan ini, maka dalam model Bayesian dengan kasus underreported (disingkat Bayesian underreported) juga akan memakai prior untuk θ adalah gamma (a, β).

$$f(\theta|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}}\theta^{\alpha-1}\exp\left(-\frac{\theta}{\beta}\right), \quad \theta \geq 0 \tag{5}$$

Selebihnya dari kasus standar, sekarang perlu ditetapkan prior untuk parameter p; yaitu peluang tecatat tidaknya seorang penderita. Mengingat p adalah sebuah peluang yang nilainya tentu dalam rentang [0,1], maka prior untuk parameter ini digunakan distribusi peluang uniform [0,1]

$$g(p) = 1, \quad 0 \leq p \leq 1 \tag{6}$$

Dalam hal sekarang besarnya pengamatan o tergantung kepada n dan distribusinya disampaikan dalam formula (3) sedangkan n tidak diketahui tetapi diketahui bentuk distribusinya; yaitu (2). Keadaan ini seperti dalam kasus Winkleman (1996). Sehingga posterior untuk *Bayesian underreported* adalah:

$$P(\boldsymbol{n},\boldsymbol{p},\boldsymbol{\theta}|\boldsymbol{o},\alpha,\beta,\boldsymbol{E}) \quad = \quad \underbrace{P(\boldsymbol{o}|\boldsymbol{n},\boldsymbol{p})}_{likelihood} \times \underbrace{P(\boldsymbol{n}|\boldsymbol{\theta},\boldsymbol{E}) \times P(\boldsymbol{\theta}|\alpha,\beta) \times P(\boldsymbol{p})}_{prior} \tag{7}$$

Dalam penulisan persamaan (5) digunakan huruf tebal untuk menyatakan vector.

Kemudian dengan mempergunakan distribusi bersyarat untuk data pengamatan tercatat O , persamaan (3) dan prior untuk parameter n, persamaan (2), dan parameter-parameter θ dan p masing-masing dalam persamaan (5) dan (6), maka persamaan posterior menjadi sebanding dengan

$$P(\mathbf{n},\mathbf{p},\boldsymbol{\theta}\,|\,\mathbf{o},\alpha,\beta,\mathbf{E}) \quad \propto \quad \prod_{i=1}^{m}(E_i\theta_i)^{n_i}\,e^{(-E_i\theta_i)} \times \frac{p_i^{o_i}(1-p_i)^{n_i-o_i}}{(n_i-o_i)!\,o_i!} \times \frac{\theta_i^{(\alpha-1)}e^{-\frac{\theta_i}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}}$$

Dipandang secara keseluruhan jelas tidak diketahui bentuk distribusi apa, walaupun komponen-komponen pembentuknya diketahui. Selain itu, dari persamaan di atas distribusi marjinal masing – masing parameter, sulit didpatkan dengan cara mengeluarkan parameter lain melalui integrasi terhadap parameter lain. Oleh karenanya upaya untuk mendapatkannya adalah melakukan sebuah simulasi untuk distribusi posterior ini dengan cara simulasi MCMC (Markov Cahin Monte Carlo).

4. PROSES MCMC

Proses simulasi distribusi posterior di atas dikerjakan melalui sampling untuk masing-amsing parameter bersyarat kepada parameter lainnya dalam distribusi tersebut; jadi dipergunakan sebuah prosedur Gibbs sampling. Prosedur sampling dikerjakan berturut-turut melalui distribusi bersyarat berikut:

$$p|n;\, n|E,\theta;\, \theta|\alpha,\beta$$

Dan dari distribusi posterior di atas, masing-masing distribusi bersyarat ini adalah sebagai berikut:

$$P(\boldsymbol{p}|\boldsymbol{n}) \propto \prod_{i=1}^{m} p_i^{o_i}(1-p_i)^{(n_i-o_i)}$$

$$P(\mathbf{n}\,|\,\mathbf{E},\boldsymbol{\theta}) \quad \propto \quad \prod_{i=1}^{m}(E_i\theta_i)^{n_i} \times \frac{(1-p_i)^{n_i}}{(n_i-o_i)!}$$

$$P(\boldsymbol{\theta}\,|\,\mathbf{E},\alpha,\beta) \quad \propto \quad \prod_{i=1}^{m} e^{(-E_i\theta_i)} \times \theta_i^{(n_i+\alpha-1)}e^{-\frac{\theta_i}{\beta}}$$

Ketiga bentuk distribusi peluang bersyarat di atas masing-masing memiliki bentuk yang diketahui. Parameter p berdistribusi distribusi beta, sehingga parameter p dapat dihasilkan melalui sampling dari distribusi beta dengan parameter (o,n-o). Formulasi distribusi peluang

bersyarat untuk *n* merupakan *kernel* dari distribusi peluang Poisson dengan nilai-nilai pengamatan tidak dimulai dari nol tetapi dari *o*, disebut sebagai *displaced Poisson distribution*, Johnson and Kotz (1970). Sedangkan parameter θ bentuk distribusi peluang bersyaratnya adalah gamma dengan parameter $(n_i + \alpha, 1/(E_i + \frac{1}{\beta}))$. Oleh karenanya masing-masing parameter ini, karena distribusi peluang bersyarat kepada parameter lain diketahui bentuknya maka dapat disimulasi memakai algorithma Gibbs Sampling.

Terkait nilai α dan β, mereka masing-masing dicari melalui maksimum likelihood dari distribusi binomial negative yang merupakan marjinal dari (α, β); yaitu

$$L(\alpha,\beta) = \frac{1}{\{\Gamma(\alpha)\}^m}\prod_{i=1}^{m}\Gamma(O_i+\alpha)\beta^{o_i}\frac{1}{\{\beta E_i+1\}^{O_i+\alpha}} \qquad (11)$$

Lawson (2000), dan solusi yang memaksimumkan likelihood ini merupakan penaksir (α, β). dicari melalui selesaikan secara numeric, lihat Clayton (1987). Penaksiran (α, β) dikerjakan di awal, sebelum simulasi untuk distribusi posterior, nilainya menjadi input dalam simulasi

Berikut adalah tahapan sampling untuk ketiga parameter:

1. Mulai dari wilayah I =1
2. Tentukan nilai awal n_i^0, p_i^0, θ_i^0
3. Untuk j=1,2, . . . melalui simulasi tentukan
   a. $n^{(j+1)}$ dari $n \mid p^j, \theta^j$
   b. $p^{(j+1)}$ dari $p \mid n^j, \theta^j$
   c. $\theta^{(j+1)}$ dari $\theta \mid p^j, n^j$
4. Wilayah = i+1,
5. stop bila seluruh m wilayah sudah disimulasi, bila belum kembali ke 2.

DAFTAR PUSTAKA

[1]. Clayton D, Kaldor J (1987). Empirical Bayes estimates of age-standardized relative risks for use in diseas e mapping. *Biometrics*; 43: 671-681.

[2]. Charnet, R., Gokhale, D.V. (2004). Statistical Inference for Damaged Poisson Distribution. *Comunication in Statistics-Simulation and Computation*, 33:2, 259-269

[3]. Lawson A B, Biggeri A B, Boehning D, Lesa!re E, VielJ-F, Bertollini R. (1999). *Disease Mapping and Risk Assessment for PublicHealth*.Wiley, NewYork.

[4]. Lawson,A.B., Biggeri, A.B., Boehning, D., LesaJre, E., Viel, J.F., Clark, A., Schlattmann, P., Divino,F., (2000). Disease mapping models: an empirical evaluation. *Statist. Med.* 19, 2217–2241.

[5]. Rao, C.R. (1965) On discrete Distribution arising out of methods of ascertainment. *Sankhya Ser A.* 27:311-324.

[6]. Rao, C.R., Rubin, H. (1964). On a characterization of the Poisson distribution. *Sankhya* 26:295-298.

[7]. Scollinik, D.P.M.(2006). A Damaged generalized Poisson model and its application to reported and under reported accident counts. *Astin Bulletin* 36(2), 463-487.

[8]. Winkelmann, R. (1996). Markov chain monte carlo analysis of under reported count dsts with application to worker absenteeism. *Empirical Economics*, 21: 575 – 587.