Statistika, Vol. 12 No. 2, 81 – 91
November 2012

Implementasi Model Poisson Bayes Berhirarki Dua-Level untuk Memodelkan Data Cacahan pada Masalah Pendugaan Area Kecil

NUSAR HAJARISMAN*, ACENG KOMARUDIN MUTAQIN, ANNEKE ISWANI AHMAD

Jurusan Statistika, Universitas Islam Bandung, Jl Purnawarman 63, Bandung, Indonesia
*Email: nrisman@yahoo.co.uk

ABSTRACT

In this paper, we address the issue of estimation of the hierarchical Bayesian models, especially for count data in small area estimation problem. This model was developed by combining the existing terminology in generalized linear models with the concept of Bayes methods, especially hierarchical Bayes methods, such that it can be implemented to address the problem of small area estimation for survey data in the form of the count data. Development of this model starts by assuming that the observed random variable is a member of the exponential family conditional on a certain parameter. The main objective of the development of this model is to make inference on these parameters are also considered as random variables. Then these parameters are modeled with the Fay-Herriot model as the basic model of the small area estimation. Furthermore, the combination of both models will be standardized in such a way as to represent a model within the framework of Bayes methods that will eventually form a two-level hierarchical Bayes Poisson model to solve problems in small area estimation.

Keywords: small area estimation, Fay-Herriot model, generalized linear models, Poisson distribution, Markov chain Monte Carlo, Gibbs sampling.

1. PENDAHULUAN

Pendugaan area kecil (*small area estimation*, SAE) merupakan teknik statistika yang tujuannya untuk memperoleh penduga pada area (*domain*) kecil, dimana penduga survey langsung tidak dapat diandalkan, bahkan kadang-kadang tidak dapat dihitung yang disebabkan oleh terbatasnya ukuran sampel yang tersedia. Pembahasan mengenai berbagai metode statistika untuk memperoleh pendugaan area kecil sudah banyak dilakukan dalam Rao (2003). Pembahasan yang paling utama adalah memperhatikan penggunaan model area kecil yang eksplisit melalui kekuatan peminjaman (*borrow strength*) dari area yang berdekatan menurut ruang atau waktu atau melalui informasi tambahan yang diperkirakan berkorelasi dengan variabel yang diamati.

Berbagai model dasar dalam pendugaan area kecil ini pada umumnya digunakan ketika variabel respons yang diamati adalah berbentuk kontinu. Padahal seringkali dalam menganalisis hubungan antara beberapa variabel, terdapat sejumlah fenomena dimana variabel responnya berbentuk data cacahan. Dalam mengamati suatu fenomena dimana variabel responnya berbentuk cacahan, maka fenomena seperti ini menyangkut banyaknya suatu kejadian yang biasanya diasumsikan mengikuti distribusi Poisson. Selanjutnya, inferensi dari penduga *model-based* ini merujuk pada suatu distribusi tertentu yang harus terpenuhi dari model yang diasumsikan, misalnya mengikuti distribusi Poisson tadi. Oleh karena itu, pemilihan model dan validasi model akan memegang peranan penting dalam pendugaan *model-based* ini. Apabila model yang diasumsikan tidak memberikan kecocokan yang baik terhadap data, maka penduga *model-based* akan menjadi bias yang pada akhirnya nanti akan membawa pada kegagalan dalam membuat inferensi yang baik.

Pada saat ini metode Bayes telah banyak digunakan untuk menangani masalah pendugaan area kecil. Metode Bayes yang sudah mulai banyak dikembangkan dalam hal ini adalah metode Bayes empirik dan Bayes berhirarki, yang secara khusus cukup baik dalam menggambarkan

hubungan sistematik dari area lokal melalui model. Namun demikian, perkembangan metode Bayes untuk masalah pendugaan area kecil saat ini masih difokuskan pada variabel yang kontinu. Padahal seringkali data yang diperoleh melalui survey ini berbentuk diskrit atau kategorik, sehingga metode Bayes empirik dan Bayes berhirarki yang dirancang untuk data kontinu menjadi tidak tepat lagi untuk diterapkan.

Pada saat data (atau dalam hal ini variabel respons) yang diamati berbentuk data diskrit atau kategorik, lebih khusus lagi berbentuk data cacahan yang berdistribusi Poisson, maka model yang dapat diterapkan adalah melalui model linear terampat (*generalized linear model*, GLM). Penggunaan pendekatan metode Bayes pada model linear terampat ini pada dasarnya sudah banyak dilakukan. Akan tetapi bagaimana penggunaan metode Bayes, khususnya metode Bayes berhirarki, pada model linear terampat yang secara langsung dapat diterapkan pada penanganan masalah pendugaan area kecil. Masih sedikit literatur yang membahas tentang masalah ini (Trevisani dan Torelli, 2007). Oleh karena itu, dalam penelitian ini akan dikembangkan model linear terampat Bayes berhirarki untuk data cacahan yang berbentuk Poisson yang akan diterapkan pada masalah pendugaan area kecil.

Model Poisson merupakan sub-bagian dari model linear terampat yang banyak digunakan, selain model logistik atau binomial. Dalam makalah ini akan dikembangkan suatu model linear terampat Bayes berhirarki yang berdistribusi Poisson. Model Bayes berhirarki yang akan dikembangkan adalah model Bayes berhirarki dua-level, atau selanjutnya model ini disebut juga sebagai model regresi Poisson Bayes berhirarki dua level yang nantinya akan diterapkan pada masalah pendugaan area kecil. Selanjutnya, isu penting yang akan dibahas dalam bab ini adalah masalah komputasi. Syarat cukup yang diberikan untuk distribusi posterior dari parameter yang diamati diharapkan akan bersifat proper di bawah model hirarki yang diusulkan ini. Prosedur Bayes yang diimplementasikan dalam penelitian ini dilakukan melalui teknik integrasi rantai Markov Monte Carlo (*Markov chain Monte Carlo*, MCMC), khususnya menggunakan teknik sampling Gibbs. Terakhir, untuk mengevaluasi performa dari model yang diusulkan akan dilakukan studi simulasi. Adapun performa model yang akan menjadi perhatian utama dari studi simulasi ini adalah ketidakbiasan dari penaksir parameter dan galat bakunya.

2. MODEL REGRESI POISSON BAYES BERHIRARKI DUA-LEVEL

Model regresi Poisson berhirarki telah banyak digunakan untuk menganalisis berbagai jenis data yang berbentuk cacahan. Kebanyakan analisis yang dilakukan untuk pemetaan penyakit (*desease mapping*) dimulai dengan proses sampling Poisson. Clayton dan Klador (1987) menggambarkan pendekatan Bayes empirik yang memperhatikan kemiripan spasial antar angka kematian penyakit tertentu. Sementara itu Bernadineli dan Montomoli (1992) membandingkan metode Bayes empirik dan Bayes berhirarki, dimana metode Bayes berhirarki ini diimplementasikan melalui metode rantai Markov Monte Carlo (MCMC). Sementara itu, Breslow dan Clayton (1993) menggunakan model campuran linear terampat untuk mempelajari masalah pemetaan penyakit ini. Sedangkan, Waller, et al. (1997) mengusulkan model Bayes berhirarki spatio-temporal untuk memodelkan angka kematian regional menurut ruang dan waktu termasuk didalamnya interaksi antara ruang dan waktu itu sendiri. Selain itu, model Bayes berhirarki dua-level digunakan oleh Nandram, Sendransk, dan Pickle (1999) untuk menduga angka kematian pada Wilayah Layanan Kesehatan Amerika Serikat. Mereka juga menggunakan model tersebut untuk pemetaan angka kematian pada penyakit *obstructive pulmonary* kronis (Nandram, Sendransk, dan Pickle (2000)).

Dalam penelitian ini akan diusulkan pengembangan model regresi Poisson berhirarki yang pertama kali diusulkan oleh Christiansen dan Morris (1997), dimana model ini pada awalnya tidak dirancang untuk digunakan dalam masalah survey sampling. Sekali lagi, model ini dikembangkan dengan cara memadukan terminologi yang ada dalam model linear terampat dengan konsep metode Bayes, khususnya metode Bayes berhirarki, sedemikian rupa sehingga dapat diimplementasikan untuk menangani masalah pendugaan area kecil untuk data survey yang berbentuk data cacahan. Pengembangan model ini dimulai dengan mengasumsikan variabel acak yang diamati merupakan anggota dari keluarga eksponensial, sebagaimana yang muncul dalam konsep pemodelan linear terampat, bersyarat pada suatu parameter tertentu. Tujuan utama dari pengembangan model ini adalah membuat inferensi pada parameter tersebut yang juga dianggap sebagai variabel acak. Kemudian parameter tersebut dimodelkan dengan menggunakan model Fay-Herriot sebagai model dasar dalam konsep pendugaan area

kecil (Fay dan Herriot, 1979). Selanjutnya, perpaduan dari kedua model tersebut akan distandarkan sedemikian rupa sehingga mewakili suatu model dalam kerangka kerja metode Bayes yang pada akhirnya akan terbentuk model Poisson Bayes berhirarki dua-level untuk menyelesaikan masalah dalam pendugaan area kecil.

Model regresi Poisson Bayes berhirarki dua-level ini dimulai dengan memisalkan y_i sebagai variabel yang menyatakan banyaknya peristiwa 'sukses' atau dalam hal ini banyaknya kejadian yang mati pada area ke-i, n_i menyatakan populasi dalam area ke-i, serta λ_i menyatakan angka mortalitas pengamatan pada area ke-i, dimana $\lambda_i = y_i/n_i$ (untuk $i = 1, 2, \ldots, m$), dan m menunjukkan banyak area kecil yang diamati. Untuk merumuskan model regresi Poisson Bayes berhirarki multi-level, dilakukan melalui tiga level, yaitu level 1 dari model deskriptif menyatakan distribusi dari vektor data yang diamati, $\mathbf{y} = (y_1, \ldots, y_m)`$ dengan syarat pada parameter individu $\{\lambda_i\}$; Level 2 untuk menyatakan distribusi gamma untuk $\{\lambda_i\}$ dengan syarat pada hyperparameter $\alpha \equiv (\tau, \boldsymbol{\beta})$; serta level 3 untuk menyatakan distribusi dari parameter terstruktur $(\tau, \boldsymbol{\beta})$.

Level 1: Model Individu

Asumsi dasar yang harus dipenuhi adalah bahwa variabel respons y_i mengikuti distribusi Poisson untuk parameter λ_i yang tetap (*fixed*), yaitu:

$$y_i|\lambda_i \sim \text{Poisson}(n_i\lambda_i), \text{ untuk } i = 1, \ldots, m \quad \ldots (1)$$

Diketahui bahwa angka mortalitas pengamatan, $\mathbf{z}$, dimana $z_i \equiv y_i/n_i$, dimana angka mortalitas pengamatan ini mempunyai nilai harapan $E(z_i) = \lambda_i$.

Level 2: Model Terstruktur

Parameter Poisson individu $(\lambda_1, \ldots, \lambda_m)$ mengikuti distribusi gamma yang bersifat *conjugate* untuk $i = 1, \ldots, m$ yang saling bebas dengan syarat pada vektor hyperparameter yang tidak diketahui $\alpha = (\tau, \beta_0, \ldots, \beta_{k-1})$, dimana k menyatakan banyaknya koefisien regresi. Dengan demikian

$$\lambda_i|\tau, \boldsymbol{\beta} \sim \text{Gamma}\left(\tau, \frac{\tau}{\mu_i}\right) = \mathbf{x}_i`\boldsymbol{\beta} \quad \ldots (2)$$

Fungsi hubung log (yang merupakan fungsi hubung alamiah untuk distribusi Poisson) diasumsikan untuk rata-rata terstruktur, sehingga $\log(\mu_i) = \mathbf{x}_i`\boldsymbol{\beta}$ untuk kovariat yang bersifat tetap (*fixed*) $\mathbf{x}_i = (x_{i0}, \ldots, x_{i,k-1})`$ dan $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{k-1})`$. Dalam hal ini parameter τ merupakan cacahan prior yang tidak teramati, dan tidak harus berupa bilangan bulat.

Perhatikan bahwa model Poisson-gamma bersifat *conjugate* yang akan membawa pada inferensi posterior mengenai parameter eksak λ_i bersyarat pada τ dan β, dimana distribusi posteriornya dapat dinyatakan sebagai berikut:

$$\lambda_i|\boldsymbol{\beta}, \tau, \mathbf{y} \sim \text{Gamma}\left(y_i + \tau, n_i + \frac{\tau}{\mu_i}\right) \quad \ldots (3)$$

Dalam penelitian ini juga akan dipertimbangkan suatu distribusi prior yang tidak harus bersifat *conjugate*. Gelman (2006) memperkenalkan suatu distribusi prior yang bersifat conditionally *conjugate*. Suatu keluarga distribusi prior p(λ) merupakan conditionally *conjugate* untuk parameter λ apabila distribusi posterior bersyarat, $p(\lambda|y)$ juga berada dalam kelas tersebut. Untuk keperluan komputasi, conditionally *conjugate* mempunyai makna bahwa apabila memungkinkan untuk mengambil λ yang berasal kelas distribusi prior tertentu, maka hal ini juga memungkinkan untuk membentuk Gibbs sampler bagi λ dalam distribusi posterior.

Menurut Gelman (2006) bahwa distribusi prior yang bersifat conditionally *conjugate* merupakan konsep yang sangat bermanfaat apabila diterapkan pada model Bayes yang berhirarki. Dalam penelitian ini akan dipertimbangkan suatu distribusi prior yang lain bagi parameter λ yang bersifat conditional *conjugate*. Distribusi prior itu adalah invers-gamma, yang dinyatakan dalam

$$\lambda_i|\tau, \boldsymbol{\beta} \sim \text{IGamma}\left(\tau, \frac{\tau}{\mu_i}\right) \quad \ldots (4)$$

Gelman (2006) menambahkan bahwa jika parameter λ mempunyai distribusi prior invers gamma, maka distribusi posterior bersyaratnya adalah

$$\lambda_i|\boldsymbol{\beta}, \tau, \mathbf{y} \sim \text{IGamma}\left(y_i + \tau, n_i + \frac{\tau}{\mu_i}\right) \quad \ldots (5)$$

yang juga berdistribusi invers gamma.

Level 3: Distribusi pada Parameter Terstruktur

Untuk hyperparameter $\boldsymbol{\beta}$ dan τ, disini akan menggunakan distribusi prior

$$\pi(\boldsymbol{\beta}, \tau) \propto \frac{z_0 \tau}{(z_0 + \tau)^2} (\tau, \beta)` \in \Re^{k+1}$$

dimana $z_0 = n_0 r_0$, dengan $n_0 = \min_i n_i$ dan $r_0 = \sum_{i=1}^{m} y_i / \sum_{i=1}^{m} n_i$. Distribusi prior pada τ bersifat proper dan dipilih sedemikian rupa sehingga penduga kemungkinan maksimum bagi τ bersifat finite. Hal ini mengakibatkan bahwa distribusi posterior bagi τ yang juga bersifat proper, namun hal ini merupakan distribusi prior yang improper bagi $\boldsymbol{\beta}$. Namun demikian, menurut Christiansen dan Morris (1997) fungsi densitas posterior bersama $\boldsymbol{\beta}$ dan τ merupakan fungsi densitas yang bersifat proper pada saat banyaknya kasus (c_0) untuk $z_i > 0$ adalah setidaknya sebesar r sehingga diperoleh submatriks $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)`$ berukuran $c_0 \times r$ yang berpangkat penuh.

3. MASALAH KOMPUTASI

Pada dasarnya akan sangat sulit untuk menghitung besaran yang sedang dikaji dalam masalah parametrik yang bersifat nonlinear, sehingga perlu dilakukan penyederhanaan pendekatan masalah komputasi yang biasa digunakan, misalnya seperti metode rantai Markov Monte Carlo (MCMC). Di sini akan dibahas mengenai metode Bayes dengan distribusi prior dua-tahap yang akan menghasilkan distribusi posterior bagi dua buah hyperparameter. Perlu diketahui bahwa metode yang saat ini berkembang biasanya tidak memperoleh distribusi posterior bersyarat dalam bentuk persamaan tertutup yang mengakibatkan sampel Gibbs (Gelfand dan Smith, 1990) agak sulit untuk digunakan. Untuk mengatasi masalah tersebut kemudian digunakan algoritma Metropolis-Hasting. Namun perlu dicatat bahwa jika distribusi bersyarat posterior nonstandar berbentuk log konkaf, maka Gibb sampling dapat digunakan dengan menggunakan algoritma Gilks-Wild (Nandram, 2000). Selain itu menurut Hobert dan Casella (1996) walaupun distribusi posterior untuk model ini sulit tersedia dalam bentuk persamaan tertutup, akan tetapi struktur *conjugate* dari spesifikasi priornya dapat digunakan untuk perhitungan yang komples dari Gibbs bersyarat. Dengan demikian sample Gibbs dapat digunakan untuk mengekplorasi distribusi posterior tanpa perlu memperhatikan sifat-sifat dari distribusi posteriornya.

Nandram (2000) menyatakan bahwa distribusi posterior bersama bagi parameter yang diamati akan bersifat proper untuk sembarang model. Dalam penelitian ini, masalah komputasi dilakukan dengan mendasarkan dengan apa yang dilakukan oleh Ghosh et al. (1998) mengenai penerapan model linear terampat pada pendugaan area kecil. Proses komputasi dilakukan pada m buah area lokal atau m strata. Misalkan Y_{ik} menyatakan statistik cukup minimal (diskrit atau kontinu) yang berhubungan dengan unit ke-k dalam strata ke-i ($k = 1, \ldots, n_i; i = 1, \ldots, m$). Variabel acak Y_{ik} diasumsikan sebagai variabel acak yang saling bebas dengan fungsi densitas peluang sebagai berikut:

$$f(y_{ik}|\theta_{ik}, \phi_{ik}) = exp\left[\phi_{ik}^{-1}\left(y_{ik}\theta_{ik} - \psi(\theta_{ik})\right) + \rho(y_{ik}; \phi_{ik})\right] \quad \ldots (6)$$

dimana $k = 1, \ldots, n_i; i = 1, \ldots, m$. Model yang diberikan dalam Persamaan (6) merupakan bentuk dari model linear terampat, sebagaimana yang ditunjukkan oleh McCullah dan Nelder (1989). Fungsi densitas yang diberikan dalam (6) diparameterisasi terhadap parameter kanonik θ_{ik} dan parameter skala $\phi_{ik} > 0$. Dalam hal ini parameter skala ϕ_{ik} diasumsikan diketahui nilainya.

Parameter alamiah θ_{ik} terlebih dahulu dimodelkan sebagai

$$h(\theta_{ik}) = \mathbf{x}_{ik}`\boldsymbol{\beta} + u_i + \varepsilon_{ik} (k = 1, \ldots, n_i; i = 1, \ldots, m) \quad \ldots (7)$$

dimana h merupakan fungsi naik; $\mathbf{x}_{ik}$ adalah vektor rancangan berukuran ($p \times 1$), $\boldsymbol{\beta}$ adalah vektor koefisien regresi berukuran ($p \times 1$), u_i merupakan efek acak, dan ε_{ik} adalah galat. Disini diasumsikan bahwa u_i dan ε_{ik} adalah saling bebas dengan $u_i \sim N(0, \sigma_u^2)$ dan $\varepsilon_{ik} \sim N(0, \sigma^2)$. Apabila diperhatikan lebih jauh, model yang diberikan dalam Persamaan (7) merupakan model Fay-Herriot yang dijadikan sebagai model dasar dalam pendugaan area kecil. Dengan demikian dapat dikatakan bahwa model linear terampat dapat dihubungkan ke masalah pendugaan area kecil melalui hubungan antara model dalam Persamaan (6) dan (7).

Apabila melihat lebih jauh persamaan yang dinyatakan dalam (6) dan (7) tidak membentuk model Bayes berhirarki. Akan tetapi model tersebut akan distandarkan sedemikian rupa sehingga mewakili suatu model dalam kerangka kerja metode Bayes sebagaimana yang telah dilakukan oleh Ghosh et al. (1998). Misalkan $R_u = \sigma_u^{-2}$ dan $R = \sigma^{-2}$. Dimisalkan bahwa $\boldsymbol{\Theta} =$

$(\theta_{11}, \dots, \theta_{1n_1}, \dots, \theta_{m1}, \dots, \theta_{mn_m})$` dan $\mathbf{u} = (u_1, \dots, u_m)$`. Kemudian model hirarki yang dipertimbangkan adalah

I. Bersyarat pada $\boldsymbol{\Theta}$, $\boldsymbol{\beta}$, $\mathbf{u}$, $\mathbf{R}_u = r_{\mathbf{u}}$ dan $R = r$, dimana variabel acak Y_{ik} adalah saling bebas dengan fungsi densitas yang diberikan dalam (6);

II. Bersyarat pada $\boldsymbol{\beta}$, $\mathbf{u}$, $\mathbf{R}_u = r_{\mathbf{u}}$ dan $R = r$, dan $h(\theta_{ik}) \sim N(\mathbf{x}_{ik}`\boldsymbol{\beta} + u_i, r^{-1})$;

III. Bersyarat pada $\boldsymbol{\beta}$, $\mathbf{R}_u = r_{\mathbf{u}}$ dan $R = r$, dan $u_i \sim N(0, r_u^{-1})$. Untuk melengkapi model Bayes berhirarki, Ghosh et al. (1998) menentukan distribusi prior untuk $\boldsymbol{\beta}$, $\mathbf{R}_u = r_{\mathbf{u}}$ dan $R = r$.

IV. Besaran $\boldsymbol{\beta}$, $\mathbf{R}_u = r_{\mathbf{u}}$ dan $R = r$ adalah saling bebas dengan $\boldsymbol{\beta} \sim$ uniform$(\mathbf{R}^p)$, untuk $p<m$, $R_u \sim$ gamma $\left(\frac{1}{2}a, \frac{1}{2}b\right)$, dan $R \sim$ gamma $\left(\frac{1}{2}c, \frac{1}{2}d\right)$.

Pada bagian (IV) suatu variabel acak $Z \sim$ gamma(α, β) apabila Z mempunyai fungsi densitas peluang sebagai berikut:

$$f(z) = \frac{\beta^{\alpha} exp(-\beta z) z^{\alpha-1}}{\Gamma(\alpha)}, \text{untuk } z > 0$$

Ketidakpastian yang dinyatakan dalam model I – IV berisi dua buah komponen, yaitu (i) efek atau pengaruh dari area lokal, dan (ii) komponen galat, yang berarti bahwa model tersebut mempertimbangkan masalah kelebihan dispersi (*overdispersion*) dengan cara menyertakan suatu komponen keragaman ekstra.

Perhatian utama dalam penelitian ini adalah menemukan distribusi posterior bagi $g(\theta_{ik})$ yang bersyarat pada data $\mathbf{y} = (\theta_{11}, \dots, \theta_{1n_1}, \dots, \theta_{m1}, \dots, \theta_{mn_m})$`, dimana g adalah fungsi naik dan secara khusus untuk menemukan rata-rata, varians dan kovarians posterior dari parameter yang berada dalam model. Dalam aplikasi tertentu diketahui bahwa $g(\theta_{ik}) = \psi(\theta_{ik}) = E(\mathbf{y}_{ik}|\theta_{ik})$.

Untuk dapat menemukan rata-rata, varians dan kovarians, maka perlu diyakinkan terlebih dahulu bahwa distribusi posterior bersama dari Θ_{ik} dengan syarat $\mathbf{y}$ adalah bersifat proper. Misalkan θ_{ik} dibatasi pada selang terbuka $(\Theta_{ik}^{lo}, \Theta_{ik}^{up})$, dimana batas bawah dari interval dapat berupa $-\infty$, batas atas dari interval dapat berupa $+\infty$. Teorema untuk mendukung masalah ini diberikan sebagai berikut: diasumsikan $a> 0$, $c> 0$, $\sum_i n_i - p + d > 0$, dan $m + b > 0$. Kemudian jika

$$\int_{\Theta_{ik}^{lo}}^{\Theta_{ik}^{lo}} exp\{[\theta_{yik} - \phi(\Theta)]/\phi_{ik}\} h`(\theta) d\theta < \infty \qquad \dots (8)$$

Untuk semua y_{ik} dan $\phi_{ik}(> 0)$, maka fungsi densitas posterior bersama bagi θ_{ik} dengan syarat $\mathbf{y}$ adalah bersifat proper.

Untuk kasus dimana variabel respons berbentuk data cacahan yang mengikuti distribusi Poisson,

$$Y_{ik}|\theta_{ik} \sim \text{Poisson}(\exp(\theta_{ik}))$$

Kemudian, jika h merupakan fungsi hubung kanonik, dan $g(\theta_{ik}) = \psi(\theta_{ik}) = exp(\theta_{ik})$, maka kondisi dalam Persamaan (8) akan menghasilkan

$$\int_0^{\infty} \tau_{ik}^{y_{ik}-1} \exp(-\tau_{ik})\, d\tau_{ik} < \infty$$

akan terpenuhi pada saat $y_{ik} = 1,2, \dots$.

Evaluasi langsung untuk distribusi posterior bersama bagi $g(\theta_{ik})$ dengan syarat y melibatkan integrasi numerik berdimensi tinggi, dan masalah ini dapat diselesaikan melalui metode rantai Markov Monte Carlo (MCMC), seperti menggunakan Gibbs sampling. Implementasi dari Gibbs sampling memerlukan sampel yang dibangkitkan dari distribusi posterior bersyarat. Misalkan

$$\mathbf{h}(\boldsymbol{\Theta}) = \left(h(\theta_{11}), \dots, h(\theta_{1n_1}), \dots, h(\theta_{m1}), \dots, h(\theta_{mn_m})\right)`$$

$$\mathbf{X} = (\mathbf{x}_{11}, \dots, \mathbf{x}_{1n_1}, \dots, \mathbf{x}_{m1}, \dots, \mathbf{x}_{mn_m})$$

Dan $\mathbf{X`X}$ adalah nonsingular. Kemudian distribusi posterior bersyarat yang diperlukan berdasarkan pada model Bayes berhirarki yang diberikan dalam (I) – (IV) adalah

i. $\boldsymbol{\beta}|\boldsymbol{\Theta}, \mathbf{u}, \mathbf{r_u}, \mathbf{r}, \mathbf{y} \sim N((\mathbf{X`X})^{-1}(\mathbf{X`h}(\boldsymbol{\Theta}) - \sum_i u_i \sum_k \mathbf{x}_{ik}), r^{-1}(\mathbf{X`X})^{-1})$;

ii. $u_i|\boldsymbol{\Theta}, \boldsymbol{\beta}, \mathbf{r_u}, \mathbf{r}, \mathbf{y} \sim N\left((rn_i + r_u)^{-1} \sum_k (h(\theta_{ik}) - \mathbf{x}_{ik}`\boldsymbol{\beta})\,; (rn_i + r_u)^{-1}\right)$;

iii. $R|\boldsymbol{\Theta}, \boldsymbol{\beta}, \mathbf{r_u}, \mathbf{u}, \mathbf{y} \sim \text{Gamma}\left(\frac{1}{2}\left(c + \sum_i \sum_k (h(\theta_{ik}) - x_{ik}`\beta - u_i)^2\right); \frac{1}{2}(d + \sum_{i=1}^m n_i)\right)$;

iv. $R_u|\boldsymbol{\Theta}, \boldsymbol{\beta}, \mathbf{r}, \mathbf{u}, \mathbf{y} \sim \text{Gamma}\left(\frac{1}{2}\left(a + \sum_i u_i^2\right); \frac{1}{2}\left(b + \sum_{i=1}^m n_i\right)\right);$

v. $\theta_{ik}|\boldsymbol{\beta}, \mathbf{u}, \mathbf{r_u}, \mathbf{r}, \mathbf{y} \sim \pi(\theta_{ik}|\boldsymbol{\beta}, \mathbf{u}, \mathbf{r_u}, \mathbf{r}, \mathbf{y}) \propto$

$$\exp\left[\left(y_{ik}\theta_{ik} - \psi(\theta_{ik})\right)\phi_{ik}^{-1} - \frac{r}{2}\left(h(\theta_{ik}) - x_{ik}^{`}\beta - u_i\right)^2\right] h^{`}(\theta_{ik})$$

Sampel dapat dibangkitkan dengan mudah dari distribusi normal dan gamma sebagaimana yang diberikan dalam (i) – (iv). Namun demikian, distribusi posterior bersyarat dalam (v), dalam hal ini parameter θ_{ik} bersyarat pada $\boldsymbol{\beta}, \mathbf{u}, \mathbf{r_u}, \mathbf{r}, \mathbf{y}$diketahui hanya pada konstanta multiplikatif, sehingga hal ini menjadi kesulitan dalam mengambil sampel dari distribusi posterior bersyarat seperti itu. Dalam kasus khusus dimana $h(z) = z$ untuk seluruh z, Ghosh et al. (1998) telah menunjukkan bahwa $\log \pi(\theta_{ik}|\boldsymbol{\beta}, \mathbf{u}, \mathbf{r_u}, \mathbf{r}, \mathbf{y})$ merupakan fungsi konkaf bagi Θ_{ik}. Dalam penelitian ini, pada Bagian (iii) dan (iv) akan dilakukan dengan berdasarkan pada ditribusi prior invers gamma, yang merupakan distribusi yang bersifat non*conjugate* bagi distribusi Poisson, yaitu:

$$R|\boldsymbol{\Theta}, \boldsymbol{\beta}, \mathbf{r_u}, \mathbf{u}, \mathbf{y} \sim \text{IGamma}\left(\frac{1}{2}\left(c + \sum_i \sum_k \left(h(\theta_{ik}) - x_{ik}^{`}\beta - u_i\right)^2\right); \frac{1}{2}\left(d + \sum_{i=1}^m n_i\right)\right);$$

dan

$$R_u|\boldsymbol{\Theta}, \boldsymbol{\beta}, \mathbf{r}, \mathbf{u}, \mathbf{y} \sim \text{IGamma}\left(\frac{1}{2}\left(a + \sum_i u_i^2\right); \frac{1}{2}\left(b + \sum_{i=1}^m n_i\right)\right);$$

Inferensi mengenai $\boldsymbol{\Theta}$ berdasarkan pada yang diberikan dalam (i) – (v) secara langsung dapat diperoleh dengan cara membentuk hasil analisis output dari Gibbs sampling. Artinya, $E(\theta_{ik}|\mathbf{y})$, $V(\theta_{ik}|\mathbf{y})$, dan $\text{cov}(\theta_{ik}, \theta_{i`k`}|\mathbf{y})(i,k) \neq (i`,k`)$ dapat dengan mudah diperoleh dari rumusan untuk nilai harapan dan varians bersyarat yang diiterasikan. Hal ini merupakan penduga Rao-Blackwell sebagaimana yang ditunjukkan oleh Gelfand dan Smith (1990).

4. HASIL STUDI SIMULASI

Pada bagian ini akan dibahas mengenai studi simulasi yang dilakukan dengan tujuan untuk mengevaluasi performa dari model Bayes berhirarki dua-level dengan variabel respons yang mengikuti distribusi Poisson. Studi simulasi ini pertama dengan menspesifikasikan proses simulasinya, proses pembangkitan data, dan pembahasan hasil-hasil dari studi simulasi. Adapun performa model yang akan menjadi perhatian utama dari studi simulasi ini adalah penaksir parameter dan galat bakunya.

Spesifikasi Simulasi

Simulasi berdasarkan model Bayes berhirarki dua-level untuk variabel respons yang mengikuti distribusi Poisson sebagaimana yang ditunjukkan pada Persamaan (1), atau dalam hal ini $y_i|\lambda_i \sim \text{Poisson}(n_i\lambda_i)$, untuk i = 1, 2, …, m. Dalam hal ini λ_i merupakan parameter mengenai angka mortalitas (*mortality rate*) yang diasumsikan mengikuti distribusi gamma, atau dapat dinyatakan dalam bentuk $\lambda_i|\tau, \beta \sim \text{Gamma}(a, b) = \mathbf{x}_i^{`}\boldsymbol{\beta}$. Perlu diketahui bahwa parameter λ_i yang berdistribusi gamma ini merupakan level pertama dari model Bayes berhirarki dua-level, sedangkan level kedua dari hirarki ini terletak pada parameter gamma a yang berdistribusi hyperprior, $h_a(\upsilon)$ dan parameter gamma b yang berdistribusi hyperprior $h_b(\rho)$, dimana ν dan ρ masing-masing menunjukkan parameter dari distribusi hyperprior tersebut.

Selanjutnya, simulasi ini dilakukan dengan spesikasi bahwa sampel diambil diambil berdasarkan pada distribusi bersyarat penuh. Untuk implementasinya simulasi ini akan dilakukan dengan spesifikasi sebagai berikut:

a. Menetapkan multipel-run secara paralel yaitu sebanyak L = 5 yang masing-masing mempunyai panjang 'burn-in' sebesar B = 1000, serta ukuran sampling Gibbs sebesar G = 5000.
b. Menetapkan nilai parameter gamma a dan b yang relatif cukup mungkin, dalam ini digunakan nilai $a = b$ = 0.002.
c. Menetapkan prior untuk hyperparameter a dan b yang masing-masing juga mengikuti distribusi gamma.

Spesifikasi simulasi kedua yang dipertimbangkan dalam penelitian ini adalah dengan menggunakan distribusi prior untuk parameter τ berdasarkan pada distribusi invers gamma atau $\lambda_i|\tau, \beta \sim \text{IGamma}(a, b)$, dengan mengambil nilai a dan b sama seperti pada spesifikasi

simulasi pertama, yaitu nilai $a = b = 0.002$. Sedangkan prior untuk hyperparameter a dan b yang masing-masing juga mengikuti distribusi invers gamma.

Pembangkitan Data

Untuk keperluan studi simulasi ini, variabel respons, y_i, dibangkitkan dengan menggunakan terminologi yang ada dalam model linear terampat. Dalam hal ini variabel respons, y_i, diasumsikan mengikuti distribusi Poisson dan menggunakan fungsi hubung log. Oleh karena dalam penelitian ini memerlukan variabel ni, yang di dalam terminologi model linear terampat dianggap sebagai variabel *exposure*, maka model yang dipertimbangkan untuk keperluan pembangkitan data adalah:

$$\log \mu_i = \eta_i = o_i + \mathbf{x}_i^T \beta \qquad \ldots (9)$$

Dalam model ini variabel respons diasumsikan sebagai $y_i \sim Poisson(\mu_i)$, dimana $\mu_i = n_i \lambda_i$. Untuk kasus dalam penelitian ini parameter λ_i merupakan angka mortalitasnya. Besaran o_i dalam Persamaan (9), dalam terminologi model linear terampat disebut juga sebagai *offset,* dan dirumuskan sebagai $o_i = \log(n_i)$. Variabel *offset* ini merupakan kovariat dalam prediktor linear dimana koefisiennya tidak ditaksir, akan tetapi diasumsikan sama dengan satu.

Dalam studi simulasi ini akan dilakukan untuk model regresi Poisson dengan dua buah variabel prediktor, x_1 dan x_2, sehingga model yang dipertimbangkan adalah

$$\log \mu_i = \eta_i = o_i + \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

Langkah-langkah proses pembangkitan data dilakukan sebagai berikut:

a. Menetapkan banyaknya pengamatan, yaitu $m = 10$.
b. Menetapkan nilai-nilai koefisien regresi, yaitu$\beta_0 = 0.5$, $\beta_1 = 1.5$, dan $\beta_2 = 2.0$.
c. Menetapkan nilai-nilai untuk variabel offset, o_i, untuk $i = 1, 2, \ldots, m$.
d. Membangkitan data untuk variabel prediktor x_1 dan x_2 dengan ketentuan:

$$x_1 \sim \text{uniform}(m, 0{,}10) \text{ dan } x_2 \sim \text{uniform}(m, 0{,}10)$$

e. Menghitung besaran rata-rata, μ_i, dengan rumus: $\mu_i = \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})$
f. Membangkitkan data untuk variabel respons yang mengikuti distribusi Poisson dengan parameter μ_i yang diperoleh pada bagian sebelumnya.

Hasil dan Pembahasan Simulasi

Hasil-hasil dari studi simulasi yang dilakukan pada penelitian ini adalah untuk melihat performa model terutama yang berkenaan ketidakbiasan dari penaksir parameter dan galat bakunya. Selain itu, inferensi Bayes berdasarkan studi simulasi ini perlu dievaluasi apakah rantai Markov telah mencapai kestasioneran dari distribusi posterior yang diinginkan atau tidak. Proses ini dilakukan melalui diagnostik kekonvergenan dengan menggunakan trace plot. Tabel 1 menyajikan ringkasan statistik berupa rata-rata dan simpangan untuk distribusi posteriornya berdasarkan pada dua penggunaan distribusi prior yang berbeda.

Tabel 1
Ringkasan Statistik untuk distribusi posterior berdasarkan prior gamma dan invers gamma

Parameter	Prior Gamma		Prior Invers Gamma	
	Rata-rata	**Simpangan Baku**	**Rata-rata**	**Simpangan Baku**
Beta0	0.5501	0.2696	0.4526	0.3355
Beta1	1.7075	0.2138	1.2536	0.6892
Beta2	1.9490	0.1093	2.1315	0.2051

Berdasarkan hasil studi simulasi yang ditunjukkan pada Tabel 1 tersebut, terlihat bahwa nilai-nilai untuk setiap parameter (β_0, β_1, dan β_2), baik yang diperoleh dari distribusi prior gamma maupun distribusi prior invers gamma memberikan nilai taksiran parameter yang tidak jauh berbeda daripada nilai parameter aslinya, dimana nilai parameter aslinya untuk keperluan pembangkitan data masing-masing adalah $\beta_0 = 0.5$, $\beta_1 = 1.5$, dan $\beta_2 = 2.0$. Namun demikian, simpangan baku untuk distribusi posterior yang dihasilkan dari distribusi prior gamma relatif

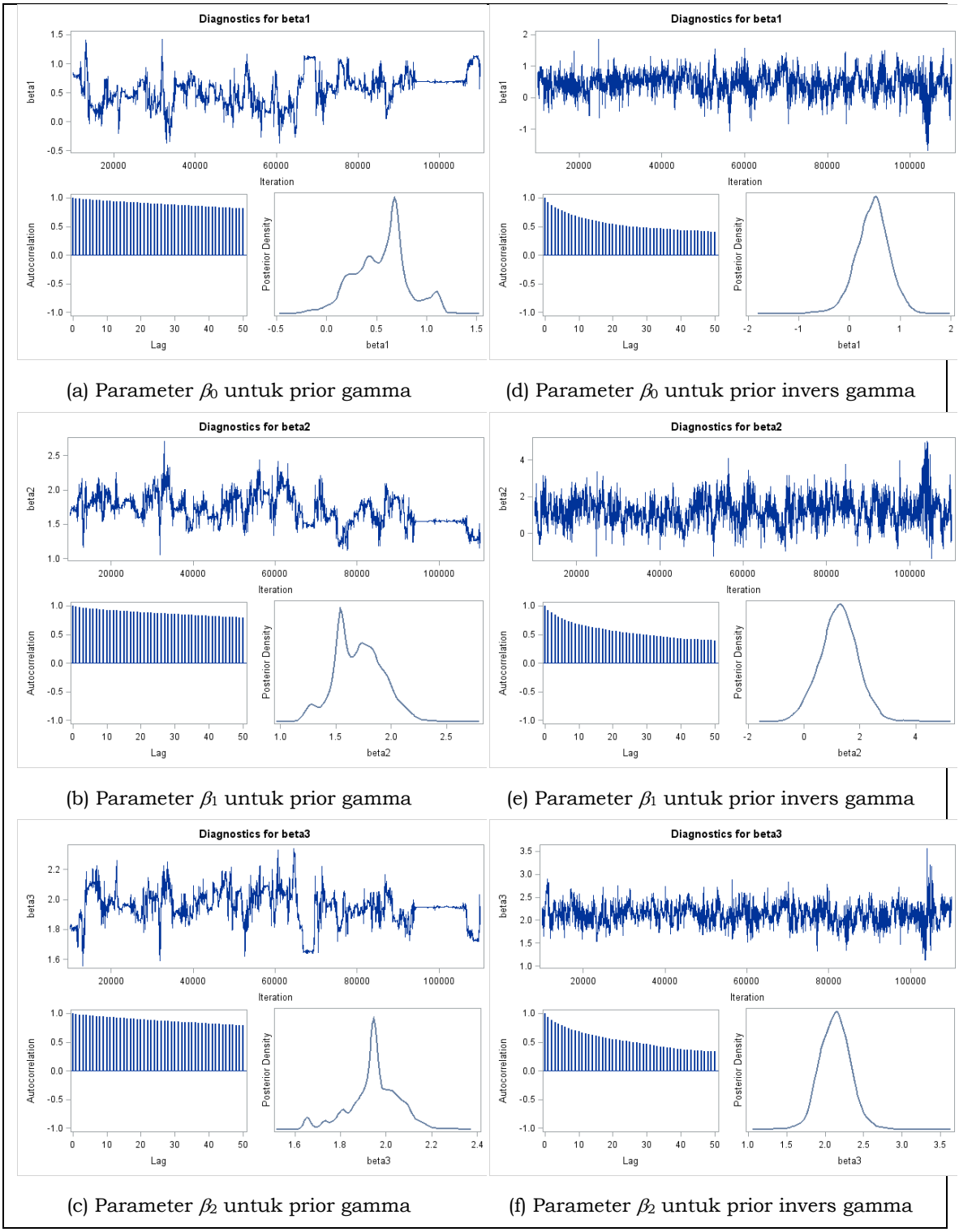
lebih kecil dibandingkan dengan simpangan baku yang dihasilkan dari distribusi prior invers gamma.

Selanjutnya, hasil dari diagnostik kekonvergenan dari studi simulasi ini dilakukan melalui trace plot sebagaimana yang ditunjukkan pada Gambar 1. Trace plot sampel dengan indeks simulasi dapat digunakan sebagai alat untuk memperkirakan apakah kekonvergenan suatu rantai Markov terhadap kestasioneran distribusinya sudah tercapai atau belum. Apabila konvergensi terhadap kestasioneran distribusi ini belum tercapai, maka perlu menambah periode 'burn-in' dalam proses simulasinya. Suatu rantai Markov dikatakan mencapai kestationeran apabila distribusi dari titik-titiknya tidak berubah sebagaimana perkembangan rantai Markovnya. Dalam hal ini dapat dilihat melalui trace plot yang relatif konstan antara rata-rata dan variansnya melalui plot fungsi otokorelasi antar sampel dan plot densitas peluang.

Berdasarkan hasil trace plot yang ditunjukkan pada Gambar 1 terlihat bahwa nilai taksiran parameter untuk setiap parameter (β_0, β_1, dan β_2) dari distribusi prior invers gamma dipusatkan di sekitar nilai yang asal yang dibangkitkan, yaitu $\beta_0 = 0.5$, $\beta_1 = 1.5$, dan $\beta_2 = 2.0$ (lihat Gambar 1d, 1e, dan 1f). Sebaliknya, nilai taksiran parameter yang berasal dari distribusi prior gamma (pada Gambar 1a, 1b, dan 1c) tidak terpusat di sekitar nilai yang asli yang dibangkitkan, bahkan hasil plotnya cenderung berfluktuasi cukup besar. Hal ini menujukkan bahwa kekonvergenan suatu rantai Markov terhadap kestasioneran distribusinya yang berasal dari distribusi prior invers gamma sudah tercapai, sedangkan untuk distribusi prior gamma belum tercapai.

Hasil di atas juga sejalan dengan plot otokorelasi antar sampel dan plot densitas peluangnya. Berdasarkan plot tersebut tampak bahwa nilai taksiran setiap parameter yang berasal dari distribusi prior menghasilkan plot otokorelasi yang penurunannya lambat, serta plot densitas peluangnya menunjukkan pola distribusi dengan lebih dari satu modus. Sementara itu, nilai taksiran setiap parameter yang berasal dari distribusi prior invers gamma cenderung menghasilkan plot otokorelasi yang penurunannya relatif lebih cepat, serta plot densitas peluang yang menghasilkan pola distribusi yang simetris.

Beradasarkan hasil dari simulasi ini, baik berdasarkan hasil ringkasan statistik maupun trace plot, menunjukkan bahwa ringkasan statistik distribusi posterior (yang ditunjukkan melalui rata-rata dan simpangan bakunya) yang berasal dari distribusi prior gamma maupun invers gamma memberikan hasil yang tidak jauh berbeda dengan nilai asli dari parameternya. Namun, berdasarkan hasil diagnostik kekonvergenan menunjukkan bahwa hasil pemodelan yang berasal dari distribusi invers gamma lebih konvergen pada kestasioneran distribusi posterior dibandingkan dengan yang berasal dari distribusi prior gamma yang pada dasarnya merupakan distribusi prior yang bersifat *conjugate* bagi distribusi Poisson. Perlu diketahui bahwa apabila proses pemodelan belum mencapai konvergen, maka hal ini dapat diatasi dengan cara memperpanjang periode 'burn-in'. Akan tetapi cara ini mempunyai kelemahan, yaitu selain proses komputasi menjadi relatif lebih lama, juga akan menghasilkan penaksir varians yang masih bersifat takbias namun penaksirnya itu bersifat overestimate terutama jika titik-titik data yang dibangkitkan mengandung masalah overdispersi.



(a) Parameter β_0 untuk prior gamma

(d) Parameter β_0 untuk prior invers gamma

(b) Parameter β_1 untuk prior gamma

(e) Parameter β_1 untuk prior invers gamma

(c) Parameter β_2 untuk prior gamma

(f) Parameter β_2 untuk prior invers gamma

Gambar 1.
Trace plot untuk studi simulasi
(a) – (c) Distribusi Prior Gamma dan (d) – (f) Distribusi Prior Invers Gamma

5. DISKUSI

Dalam penelitian ini telah ditunjukkan implementasi dari konsep pemodelan linear terampat pada masalah pendugaan area kecil. Model yang dikembangkan dalam penelitian ini adalah model regresi Poisson berhirarki dua-level dengan cara memadukan terminologi yang ada dalam model linear terampat dengan konsep metode Bayes berhirarki dua-level sedemikian rupa sehingga dapat diimplementasikan untuk menangani masalah pendugaan area kecil yang diwakili oleh model Fay-Herriot. Sebagaimana yang telah diketahui bahwa salah satu permasalahan dalam pemodelan Bayes berhirarki adalah pemilihan distribusi prior. Apabila distribusi prior ini diketahui maka inferensi dapat dengan mudah dilakukan dengan cara meminimumkan galat posterior, menghitung daerah kepekatan distribusi posterior yang lebih tinggi dimensinya, atau mengintegrasi parameter untuk menemukan distribusi prediktifnya.

Model Poisson Bayes berhirarki dua-level yang dikembangkan di sini menggunakan dua distribusi prior yang berbeda. Pertama, menggunakan distribusi prior gamma yang merupakan distribusi prior yang bersifat *conjugate* bagi distribusi Poisson. Dengan menggunakan distribusi prior yang bersifat *conjugate* ini akan memberikan kemudahan dalam penentuan distribusi posteriornya. Kedua, disini menggunakan distribusi prior yang bersifat non *conjugate* bagi distribusi Poisson, yaitu distribusi invers gamma. Penggunaan distribusi prior yang bersifat non *conjugate* ini tentu saja akan memberikan kesulitan dalam menemukan distribusi posterior dalam bentuk persamaan tertutup. Namun masalah ini dapat diatasi melalui penerapan rantai Markov Monte Carlo yang akan menghasilkan sederet variabel acak yang mendekati distribusi posteriornya.

Dalam penelitian ini juga telah diperkenalkan suatu distribusi prior yang bersifat conditionally *conjugate*. Suatu keluarga distribusi prior $p(\lambda)$ merupakan conditionally *conjugate* untuk parameter λ apabila distribusi posterior bersyarat, $p(\lambda|y)$ juga berada dalam kelas tersebut. Untuk keperluan komputasi, conditionally *conjugate* mempunyai makna bahwa apabila memungkinkan untuk mengambil λ yang berasal kelas distribusi prior tertentu, maka hal ini juga memungkinkan untuk membentuk Gibbs sampler bagi λ dalam distribusi posterior. Distribusi prior yang bersifat conditionally *conjugate* merupakan konsep yang sangat bermanfaat apabila diterapkan pada model Bayes yang berhirarki. Hasil-hasil studi simulasi menunjukkan bahwa peforma model yang berasal dari dsitribusi prior invers gamma (yang bersifat non *conjugate*) relatif tidak berbeda dengan yang berasal distribusi yang bersifat *conjugate* (distribusi gamma), bahkan telah mencapai kekonvergenan yang diinginkan. Namun demikian, sebagai tantangan ke depan terutama dalam pemodelan Bayes berhirarki, menurut Carlin dan Gelfand (1991) pemilihan prior untuk level dua atau yang lebih tinggi tidak memberikan efek yang besar terhadap hasil pemodelan. Akan tetapi apabila hal ini diterapkan pada level pertama tentu saja memberikan efek yang berarti, sehingga perlu melakukan analisis Bayes yang bersifat *robust*.

DAFTAR PUSTAKA

[1] Bernardinelli, L. and Montomoli, C. (1992). Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. *Statistics in Medicine*, **11**, 983-1007.

[2] Breslow, N.E., dan Clayton, D.G. (1993) Approximate Inference in Generalized Linear Mixed Models.*Journal of the American Statistical Association*, **88**, 9-25

[3] Christiansen, C. L. and Morris, C. N. (1997). Hierarchical Poisson regression modeling. *Journal of the American Statistical Association*, **92**, 618-632.

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

[5] Datta, G.S., Rao, J.N.K., and Smith, D.D. (2005). On Measuring The Variability of Small Area Estimators Under a Basic Area Level Model.*Biometrika*, **92**, 183-196.

[6] Fay, R.E., dan Herriot, R. (1979). Estimates of Income for Small Places: An Application of James-Stein Procedures to Cencus Data.*Journal of the American Statistical Association*, **74**, 269-277.

[7] Gelfand, A.E., dan Smith, A.F.M. (1990). Sampling-Based Approaches to Calculating Marginal Densities.*Journal of the American Statistical Association*, **85**, 398-409.

[8] Ghosh, M., Natarajan, K., Stroud, T.W.F., and Carlin, B.P. (1998). Generalized Linear Models for Small Area Estimation.*Journal of the American Statistical Association*, **93**, 273-282.

[9] Hobert, J.P., and Casella, G. (1996). The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models.*Journal of the American Statistical Association*, **91**, 1461-1473.

[10] McCullah, P. dan Nelder, J.A. (1989) *Generalized Linear Models*. Second Edition. London: Chapman and Hall.

[11] Nandram, B. (2000) Bayesian generalized linear models for inference about small area. In: Dey, D.K., Ghosh, S.K., and Mallick, B.K. (Eds.) *Generalized Linear Models: A Bayesian Perspective*. New York: Marcel Dekker, pp. 91-114.
[12] Nandram, B., Sendransk, J., and Pickle, L. (1999). Bayesian Analysis of Mortality Rates For U.S. Health Service Areas.*Sankhya: The Indian Journal of Statistics*, **61**, 146-165.
[13] Nandram, B., Sendransk, J., and Pickle, L.W. (2000). Bayesian Analysis and Mapping of Mortality Rates for Chronic Obstructive Pulmonary Disease.*Journal of the American Statistical Association*, **95**, 1110-1118.
[14] Rao, J.N.K. (2003) *Small Area Estimation*. New York: Wiley.
[15] Trevisani M, and Torelli, N. (2007). *Hierarchical Bayesian models for small area estimation with count data*. Working Paper: Dipartimento di Scienze Economiche e Statistiche, Università degli Studi di Trieste, Trieste, Italy.
[16] Waller, L., Carlin, B., Xia, H., and Gelfand, A. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, **92**, 607-617.