JOURNAL OF FUNDAMENTAL MATHEMATICS
AND APPLICATIONS (JFMA), Vol. 1, No. 2 (Nov 2018)
Available online at www.jfma.matematika.undip.ac.id

MODEL REGRESI LINIER BAYESIAN DENGAN APLIKASI PADA DATA PENUNDAAN PENERBANGAN

Vemmie Nastiti Lestari[1], Subanar[2]

[1,2]*Departemen Matematika, Universitas Gadjah Mada, Yogyakarta*
Email : [1]vemmie.lestari@ugm.ac.id , [2] subanar@ugm.ac.id

Abstract. Bayesian linear regression is an approach to linear regression where statistical analysis depend of Bayesian inference. The Bayesian model on big data uses a summary of data statistics as input; Statistical summary can be calculated from each subset, then a statistical summary of the full dataset is obtained from the sum of the summary statistics for each subset. Recent developments in data science and research, produce large datasets that are too large to be analyzed as a whole due to the limitations of computer memory or storage capacity. To overcome this, a program package was introduced from R namely `BayesSummaryStatLM` for the Bayesian linear regression model with the Markov Chain Monte Carlo implementation that overcomes this limitation. Then the program package from R, `ff` is used to read data in large datasets while calculating statistics summary. In this study Bayesian linear regression model used with several choices of prior distribution for unknown model parameters, and illustrates in simulation data and real datasets for flight delay data in US 2008. The application of simulation data and flight delay data produces a plot of density functions for the β parameters has a shape resembling a plot of Normal distribution density function, whereas for plot σ^2 parameters the density function has a shape resembling the plot of Inverse Gamma distribution density function. In the simulation data, the estimator for each parameter produced has a value that approach to the value of the specified parameter (True Value). This is also indicated by the narrow credible interval for each parameters.

Keywords: Big Data, Bayesian Method, Markov Chain Monte Carlo, Bayesian Linear Regression, BayesSummaryStatLM**.**

Abstrak. Model Regresi linier Bayesian merupakan pendekatan untuk regresi linier dimana analisis statistik yang dilakukan dalam konteks inferensi Bayesian. Perkembangan terbaru dalam ilmu data dan penelitian, menghasilkan dataset besar yang terlalu besar untuk dianalisis secara keseluruhan karena keterbatasan memori komputer atau kapasitas penyimpanan. Untuk mengatasi hal tersebut diperkenalkan paket program dari R yaitu `BayesSummaryStatLM` untuk model regresi linier Bayesian dengan implementasi Markov Chain Monte Carlo yang mengatasi keterbatasan ini. Selanjutnya paket program dari R yaitu `ff` digunakan untuk membaca data pada dataset besar sekaligus menghitung ringkasan statistik. Dalam penelitian ini digunakan model regresi linier Bayesian dengan beberapa pilihan distribusi prior untuk parameter model yang tidak diketahui, dan mengilustrasikannya pada data simulasi dan dataset real yaitu data penundaan penerbangan di US tahun 2008. Penerapan pada data simulasi maupun data penundaan penerbangan menghasilkan plot fungsi densitas untuk parameter β memiliki bentuk menyerupai plot

fungsi densitas distribusi Normal, sedangkan untuk parameter σ^2 plot fungsi densitasnya memiliki bentuk menyerupai plot fungsi densitas distribusi Inverse Gamma. Pada data simulasi, penduga untuk masing-masing parameter yang dihasilkan mempunyai nilai yang mendekati nilai parameter yang ditentukan (True Value). Hal ini juga ditunjukkan oleh sempitnya interval kredibel untuk masing-masing parameter.

Kata Kunci : Big Data, Metode Bayesian, Markov Chain Monte Carlo, Regresi Linear Bayesian, BayesSummaryStatLM**.**

I. PENDAHULUAN

Dalam era big data sekarang ini, statistisi menghadapi tantangan baru karena pertumbuhan informasi yang dihasilkan. Dalam hal ini big data mengacu pada data set yang terlalu besar dan kompleks jika menggunakan metode analisis klasik. Salah satu kesulitan utama dalam menganalisis dataset besar ini adalah pembatasan ukuran file yang dapat dibaca ke dalam memori komputer (RAM). Disamping itu, untuk dataset besar ini mungkin perlu untuk menyimpan dan mengolah dataset pada lebih dari satu mesin. Dalam penelitian ini akan dibahas Markov chain Monte Carlo (MCMC) yang diterapkan pada Model regresi linear Bayesian dengan error berdistribusi normal yang meggunakan ringkasan statistik sebagai input. Selanjutnya ringkasan statistik tersebut dapat digabungkan dari masing-masing subset. Perkembangan terbaru dalam ilmu data dan penelitian, menghasilkan dataset besar yang terlalu besar untuk dianalisis secara keseluruhan karena keterbatasan memori komputer atau kapasitas penyimpanan. Metode regresi linear Bayesian diimplementasikan dengan menggunakan bantuan paket program dari R yaitu BayesSummaryStatLM yang tersedia dari Comprehensive R Archive Network di http://CRAN.R-project.org/ [1]. Selanjutnya paket program dari R yaitu ff digunakan untuk membaca data pada dataset besar sekaligus menghitung ringkasan statistik pada masing-masing subset [2]. Ringkasan statistik pada full dataset diperoleh dari penjumlahan ringkasan statistik pada masing-masing subset.

Metode serupa yang menggunakan ringkasan statistik untuk model regresi linear Bayesian dikembangkan oleh Ordonez et al. [3] pada dataset besar. Ordonez dkk mengenalkan prosedur sistem manajemen basis data (DBMS) untuk mendapatkan ringkasan statistik untuk full dataset, kemudian melakukan regresi linier Bayesian pada ringkasan statistik tersebut. Ghosh dan Reiter [4] mengembangkan metode untuk regresi linier Bayesian berdasarkan ringkasan statistik dari subset, namun tidak dalam konteks dataset yang besar. Menurut Ordonez dkk [3] dan Ghosh dan Reiter [4] hanya mencakup satu pilihan distribusi prior untuk parameter model yang tidak diketahui dan tidak menyediakan software untuk menerapkan metode-metode tersebut. Metode yang dikembangkan oleh Evgeny Savel'ev dkk [9] yaitu regresi linier Bayesian berdasarkan ringkasan statistik dari subset untuk dataset yang besar dan *package* pada software R untuk mengakomodasi penerapan metode tersebut. Dalam penelitian ini akan digunakan software R dengan paket program untuk model regresi linier Bayesian pada dataset besar yang mencakup beberapa pilihan distribusi prior dengan parameter model yang tidak diketahui [7]. Selain itu, beberapa keunggulan fitur-fitur yang ada pada paket program diantaranya dapat mengakomodasi data yang berada di file terpisah atau yang didistribusikan melalui jaringan dan dapat menganalisis data yang diperbarui dari waktu ke waktu misalnya, *streamed* data.

Dalam penelitian ini digunakan model regresi linier Bayesian dengan beberapa pilihan distribusi prior untuk parameter model yang tidak diketahui, dan mengilustrasikannya pada data simulasi dan dataset real yaitu data penundaan penerbangan di US tahun 2008.

JOURNAL OF FUNDAMENTAL MATHEMATICS AND APPLICATIONS (JFMA), Vol. 1, No. 2 (Nov 2018)
Available online at www.jfma.matematika.undip.ac.id

II. REGRESI LINIER BAYESIAN

2.1 Regresi Linier Bayesian

Dalam statistik, regresi linier Bayesian merupakan pendekatan untuk regresi linier dimana analisis statistik yang dilakukan dalam konteks inferensi Bayesian. Saat model regresi memiliki error yang berdistribusi normal, dan jika bentuk khusus dari distribusi prior diasumsikan, hasil eksplisit tersedia untuk distribusi probabilitas posterior dari parameter model.

Model regresi linear dengan variabel respon $\mathbf{Y}$ dengan variabel prediktor $\mathbf{X}=(\mathbf{X}_1,\cdots$. Menurut [5,6] model regresi linear dengan error normal seperti berikut :

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i \quad (1)$$

dengan k merupakan variabel prediktor; $i:1,\cdots,n$

$$\varepsilon_i \sim Normal\left(0,\sigma^2\right) \quad (2)$$

Fungsi likelihood diberikan sebagai berikut :

$$p\left(\mathbf{Y}|\beta_0,\beta_1,\beta_2\cdots\beta_k,\sigma^2\right)=\frac{1}{\left(2\pi\sigma^2\right)^{n/2}}\exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\right)^T\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\right)\right] \quad (3)$$

dengan :

$\mathbf{Y}$: vektor kolom berukuran $\mathrm{n}\times 1$,

$\mathbf{X}$: matriks berukuran $\mathrm{n}\times(\mathrm{k}+1)$,

$\boldsymbol{\beta}$: vektor kolom berukuran $(\mathrm{k}+1)\times 1$

Parameter model yang tidak diketahui pada regresi linier adalah koefisien $\boldsymbol{\beta}=\left(\beta_0,\beta_1,\beta_2\cdots\beta_k\right)'$ dan parameter variansi error σ^2. Di dalam kerangka Bayesian, terlebih dahulu menetapkan distribusi prior untuk $\boldsymbol{\beta}$ dan σ^2, selanjutnya menghitung distribusi posterior bersama yang diperoleh dari fungsi likelihood dan distribusi prior dengan diasumsikan distribusi prior independen untuk $\boldsymbol{\beta}$ dan σ^2 [11]. Distribusi posterior bersyarat penuh (*full conditional posterior distributions*) untuk parameter model yang tidak diketahui sebanding dengan distribusi posterior bersama yang mana menganggap semua parameter lainnya sebagai konstanta tetap. Untuk pilihan distribusi prior untuk $\boldsymbol{\beta}$ dan σ^2 pada paket program *R*, distribusi posterior bersyarat penuh tergantung pada data melalui ringkasan statistik $\mathbf{X}^{\mathrm{T}}\mathbf{X},\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ untuk $\boldsymbol{\beta}$ dan $\mathbf{X}^{\mathrm{T}}\mathbf{X},\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ dan $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$ $\left(\mathbf{Y}^{\mathrm{T}}\mathbf{X}=\left(\mathbf{X}^{\mathrm{T}}\mathbf{Y}\right)^T\right)$ untuk σ^2.

2.2 RINGKASAN STATISTIK UNTUK DISTRIBUSI POSTERIOR BERSYARAT PENUH

Seperti yang telah dijelaskan pada sub bab sebelumnya, untuk masing-masing distribusi prior untuk $\boldsymbol{\beta}$, distribusi posterior bersyarat penuh hanya tergantung pada dua ringkasan statistik yaitu $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ dan $\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ [7]. Demikian halnya dengan distribusi posterior bersyarat penuh untuk σ^2 hanya tergantung pada tiga ringkasan statistik yaitu $\mathbf{X}^{\mathrm{T}}\mathbf{X},\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ dan $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$. Ringkasan statistik pada full dataset diperoleh dari penjumlahan ringkasan statistik pada masing-masing subset.

Di dalam paket program R yang akan digunakan, diasumsikan data dipartisi secara horizontal oleh sampel n menjadi sebanyak M subset, jika $\mathbf{X}$ mempunyai dimensi $n x \psi$, partisi tersebut dapat dituliskan sebagai berikut :

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_m \end{pmatrix} \tag{4}$$

dimana masing-masing $\mathbf{X}_m, m = 1, 2, \cdots, M$ mempunyai ψ kolom. Hal ini serupa dengan vektor $\mathbf{Y}$ juga dipartisi secara horizontal dengan partisi seperti persamaan (4). Ringkasan statistik untuk full data dapat dituliskan sebagai berikut dengan $m = 1, 2, \cdots, M$:

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} \text{ untuk full data} = \sum_{m=1}^{M} \mathbf{X}_m{}^{\mathrm{T}}\mathbf{X}_m \tag{5}$$

$$\mathbf{X}^{\mathrm{T}}\mathbf{Y} \text{ untuk full data} = \sum_{m=1}^{M} \mathbf{X}_m{}^{\mathrm{T}}\mathbf{Y}_m \tag{6}$$

$$\mathbf{Y}^{\mathrm{T}}\mathbf{Y} \text{ untuk full data} = \sum_{m=1}^{M} \mathbf{Y}_m{}^{\mathrm{T}}\mathbf{Y}_m \tag{7}$$

1. Distribusi Prior dan Distribusi Posterior Bersyarat Penuh untuk $\boldsymbol{\beta}$ dan σ^2

Untuk masing-masing distribusi prior untuk $\boldsymbol{\beta}$, distribusi posterior bersyarat penuh hanya tergantung pada dua ringkasan statistik yaitu $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ dan $\mathbf{X}^{\mathrm{T}}\mathbf{Y}$. Demikian halnya dengan distribusi posterior bersyarat penuh untuk σ^2 hanya tergantung pada tiga ringkasan statistik yaitu $\mathbf{X}^{\mathrm{T}}\mathbf{X}, \mathbf{X}^{\mathrm{T}}\mathbf{Y}$ dan $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$ dimana $\left(\mathbf{Y}^{\mathrm{T}}\mathbf{X} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{Y}\right)^{T}\right)$.

2. Distribusi Prior dan Distribusi Posterior Bersyarat Penuh untuk $\boldsymbol{\beta}$

a. Prior Uniform untuk $\boldsymbol{\beta}$

Jika diketahui distribusi prior untuk $\boldsymbol{\beta}$, dengan $\boldsymbol{\beta}$ berdimensi (k+1) x 1 yaitu $\boldsymbol{\beta} \sim$ Uniform maka distribusi posterior bersyarat penuh untuk $\boldsymbol{\beta}$ adalah (lihat [5,6])

$$\boldsymbol{\beta} \left| \sigma^2, \mathbf{X}, \mathbf{Y} \right. \sim Normal_{(k+1)}\left(\text{mean} = \left(\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\right), \text{covariance} = \left(\sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\right)\right.$$

b. Prior Normal Multivariate untuk $\boldsymbol{\beta}$ dengan vektor mean dan matrik kovariansi diketahui

Jika diketahui distribusi prior untuk $\boldsymbol{\beta}$, $\boldsymbol{\beta} \sim Normal_{(k+1)}\left(\text{mean} = \boldsymbol{\mu}\right), \text{covariance} = \left(\mathbf{C}\right)$, dimana $\boldsymbol{\mu}$ merupakan vektor berdimensi (k+1) x 1 yang ditentukan oleh peneliti dan $\mathbf{C}$ merupakan matriks simetri berdimensi (k+1) x (k+1) dan positif definit yang ditentukan oleh peneliti, maka distribusi posterior bersyarat penuh untuk $\boldsymbol{\beta}$ adalah (lihat [5,6])

$$\boldsymbol{\beta} \left| \sigma^2, \mathbf{X}, \mathbf{Y} \right. \sim Normal_{(k+1)}\left(\text{mean} = \left(\mathbf{C}^{-1} + \sigma^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\left(\mathbf{C}^{-1}\mu + \sigma^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\right)\right.,$$

$$\text{covariance} = \left(\left(\mathbf{C}^{-1} + \sigma^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\right)$$

c. Prior Normal Multivariate untuk $\boldsymbol{\beta}$ dengan vektor mean dan matrik kovariansi tidak diketahui

Jika diketahui distribusi prior untuk $\boldsymbol{\beta}$,

$$\boldsymbol{\beta}\,|\boldsymbol{\mu},\mathbf{C} \sim Normal_{(k+1)}\left(\text{mean} = \boldsymbol{\mu}\right), \text{covariance} = \left(\mathbf{C}\right),$$

dimana $\boldsymbol{\mu}$ merupakan vektor berdimensi (k+1) x 1 yang tidak diketahui, $\mathbf{C}$ merupakan matriks simetri berdimensi (k+1) x (k+1) dan positif definit yang tidak diketahui, maka distribusi hyperprior untuk $\boldsymbol{\mu}$ adalah

$$\boldsymbol{\mu} \sim Normal_{(k+1)}\left(\text{mean} = \boldsymbol{\eta}\right), \text{covariance} = \left(\mathbf{D}\right)$$

dimana $\boldsymbol{\eta}$ merupakan vektor berdimensi (k+1) x 1 yang ditentukan oleh peneliti dan $\mathbf{D}^{-1}$ merupakan matriks presisi simetri berdimensi (k+1) x (k+1) dan positif definit yang ditentukan oleh peneliti. Distribusi hyperprior untuk matriks presisi $\mathbf{C}^{-1}$ adalah

$$\mathbf{C}^{-1} \sim Wishart_{(k+1)}\left(\text{df} = \lambda,\ \text{matriks skala} = \mathbf{V}\right)$$

dimana λ merupakan derajat bebas yang ditentukan oleh peneliti dan $\mathbf{V}^{-1}$ merupakan matriks skala invers simetri berdimensi (k+1) x (k+1) dan positif definit yang ditentukan oleh peneliti.

Distribusi posterior bersyarat penuh untuk $\boldsymbol{\beta}$, $\boldsymbol{\mu}$, $\mathbf{C}^{-1}$ (lihat [5,6]), dimana untuk parameter model $\boldsymbol{\mu}$ dan $\mathbf{C}^{-1}$ yang tidak diketahui maka distribusi posterior bersyarat penuhnya tidak tergantung pada ringkasan statistik :

$$\boldsymbol{\beta}\,|\sigma^2,\boldsymbol{\mu},\mathbf{C}^{-1},\mathbf{X},\mathbf{Y} \sim Normal_{(k+1)}\left(\text{mean} = \left(\mathbf{C}^{-1} + \sigma^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\left(\mathbf{C}^{-1}\boldsymbol{\mu} + \sigma^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\right)\right),$$
$$\text{covariance} = \left(\left(\mathbf{C}^{-1} + \sigma^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\right)$$
$$\boldsymbol{\mu}\,|\boldsymbol{\beta},\sigma^2,\mathbf{C}^{-1},\mathbf{X},\mathbf{Y} \sim Normal_{(k+1)}\left(\text{mean} = \left(\mathbf{D}^{-1} + \mathbf{C}^{-1}\right)^{-1}\left(\mathbf{C}^{-1}\boldsymbol{\beta} + \mathbf{D}^{-1}\boldsymbol{\eta}\right)\right),$$
$$\text{covariance} = \left(\left(\mathbf{D}^{-1} + \mathbf{C}^{-1}\right)^{-1}\right)$$

$$\mathbf{C}^{-1}\,|\boldsymbol{\beta},\sigma^2,\boldsymbol{\mu},\mathbf{X},\mathbf{Y} \sim Wishart_{(k+1)}\left(\text{df} = (1+\lambda)\right),$$
$$\text{matriks skala} = \left(\left(\mathbf{V}^{-1} + (\boldsymbol{\beta}-\boldsymbol{\mu})(\boldsymbol{\beta}-\boldsymbol{\mu})^{T}\right)^{-1}\right)$$

3. Distribusi Prior dan Distribusi Posterior Bersyarat Penuh untuk σ^2

a. Prior Invers Gamma utk σ^2 dengan parameter *shape* dan *scale* diketahui

Distribusi prior untuk σ^2 adalah

$\sigma^2 \sim \text{Inverse Gamma}(a,b)$, dengan *a* dan *b* tidak diketahui.

Distribusi posterior bersyarat penuh untuk σ^2 adalah (lihat [5,6])

$$\sigma^2|\boldsymbol{\beta},\mathbf{X},\mathbf{Y} \sim Inverse\,Gamma\left(\frac{n}{2}+a,\left[\frac{1}{2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^{T}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})+\frac{1}{b}\right]^{-1}\right).$$

b. Prior *Inverse sigma squared* utk σ^2

Distribusi prior khusus untuk σ^2 , dengan prior Jeffreys untuk σ^2 [5] adalah

$\sigma^2 \sim 1/\sigma^2, \sigma^2 > 0..$

Distribusi posterior bersyarat penuh untuk σ^2 adalah (lihat [5,6])

$$\sigma^2|\boldsymbol{\beta},\mathbf{X},\mathbf{Y} \sim Inverse\,Gamma\left(\frac{n}{2},\left[\frac{1}{2}\left(\mathbf{Y}^T\mathbf{Y}-\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y}-\mathbf{Y}^T\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\right)\right]^{-1}\right),$$

2.3 Algoritma Gibbs Sampler

Setelah distribusi posterior bersyarat penuh ditentukan, Gibbs sampler dilakukan pada software R dimana langkah pertama yang dilakukan adalah memilih nilai awal secara random untuk parameter model yang tidak diketahui (2, ...,*m*), tetapi bukan untuk parameter yang pertama. Selanjutnya diberi label sebagai iterasi t = 0, sedangkan untuk jumlah sampel MCMC *t*, *t* = 1, ..., *T*, dapat dilakukan dengan langkah-langkah berikut :

a. Update parameter model pertama yang tidak diketahui dengan mengambil sampel dari distribusi posterior bersyarat penuh, dikondisikan pada nilai iterasi (*t* - 1) dari parameter model yang tidak diketahui.
b. Update masing-masing parameter model yang tidak diketahui (2, ...,*m*) dengan sampling dari distribusi posterior bersyarat penuh, dikondisikan pada parameter yang tersisa, sehingga akan menjadi nilai iterasi (*t*) untuk parameter yang telah diupdate, dan iterasi (*t* - 1) untuk parameter yang belum diupdate.

Dalam software *R*, yang pertama dilakukan adalah update $\boldsymbol{\beta}$ sehingga vektor tersebut tidak memerlukan nilai awal. Nilai output dari paket program *R* adalah sampel *T* dari distribusi posterior bersyarat penuh dengan parameter model yang tidak diketahui. *m*-tupel sampel dari semua parameter yang diperoleh pada iterasi (t) konvergen dalam distribusi ke hasil imbang dari distribusi posterior bersama yang benar [5].

III. PENERAPAN DATA

3.1 Data Simulasi

Prosedur data simulasi pada software R dengan paket `BayesSummaryStatLM` dalam model regresi linear Bayesian dengan data yang tersedia pada paket tersebut. Data yang digunakan adalah data simulasi yang terdiri dari satu variabel respon dan lima variabel prediktor $(y, x_1, x_2, x_3, x_4, x_5)$ dengan jumlah observasi sebanyak 50.000. masing-masing variabel prediktor disimulasikan dari distribusi normal multivariate dan diasumsikan masing-masing prediktor saling independen,

$$\boldsymbol{X} \sim Normal(0, \boldsymbol{\Sigma})$$

Matriks varian-kovariansi $\boldsymbol{\Sigma}$ didefinisikan :

$$\Sigma_{hh} = 1, h = 1, \cdots \qquad = \rho, h \neq h'$$

Dalam hal ini diambil $\rho = 0{,}2$ sehingga masing-masing prediktor mempunyai tingkat korelasi moderat cenderung saling independen karena ρ diambil yang mendekati nol. Model parameter $\boldsymbol{\beta}$ disimulasikan dari distribusi normal standar, sedangkan variabel respon $y_i, i = 1, \cdots$ disimulasikan dari model regresi linear :

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i$$

$$\varepsilon_i \sim Normal(0, \sigma^2)$$

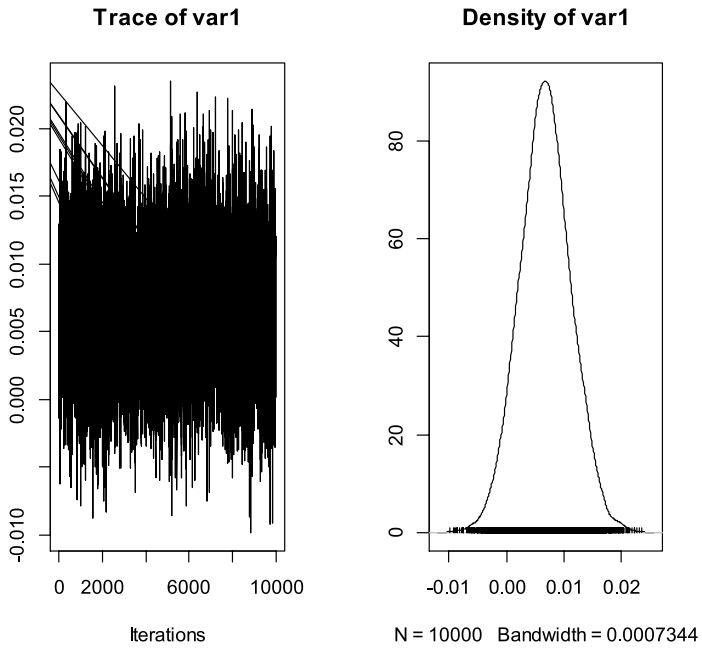
ε_i disimulasikan dari distribusi normal dengan mean = 0 dan variansi = 1.

Analisis sampel posterior MCMC dapat dilihat dari plot densitas untuk distribusi posterior marginal pada masing-masing β, ringkasan statistik untuk distribusi posterior marginal β dapat dilihat secara ringkas pada Tabel 1.

Tabel 1. Prediksi nilai posterior mean dan kuantil posterior untuk parameter model yang tidak diketahui pada data simulasi.

Parameter	True Value (Simulated)	Posterior Mean	Posterior 2,5%	Posterior 25%	Posterior 50%	Posterior 75%	Posterior 97,5%
β_0	0, 00096	0,0068	-0,0019	0,0038	0,0068	0,0097	0,0157
β_1	-1,0633	-1,061	-1,070	-1,064	-1,061	-1,058	-1,051
β_2	0,4672	0,4634	0,4543	0,4604	0,4634	0,4665	0,4727
β_3	0,7374	0,7344	0,7251	0,7311	0,7343	0,7375	0,7437
β_4	-0,2731	-0,2683	-0,2775	-0,2715	-0,2682	-0,2650	-0,2589
β_5	-0,3135	-0,3114	-0,3207	-0,3146	-0,3114	-0,3083	-0,3023
σ^2	1,00	0,9954	0,9829	0,9911	0,9955	0,9997	1,0079

Dari tabel 1. menunjukkan bahwa penduga dari masing-masing parameter dapat dilihat pada kolom posterior mean. Selanjutnya kuantil posterior 2,5% dan 97,5% menunjukkan batas bawah dan batas atas dari interval kredibel (interval kepercayaan Bayesian) untuk masing-masing parameter. Penduga untuk masing-masing parameter yang dihasilkan mempunyai nilai yang mendekati nilai parameter yang ditentukan (True Value). Hal ini juga ditunjukkan oleh sempitnya interval kredibel untuk masing-masing parameter.



Gambar 1. Iterasi Gibbs Sampler dan Plot Fungsi Densitas untuk β_0 pada data simulasi

Gambar 1. menunjukkan iterasi Gibbs sampler pada β_0, selanjutnya plot fungsi densitas untuk parameter β_0 memiliki bentuk menyerupai plot fungsi densitas distribusi Normal,

demikian seterusnya sampai β_5. Sedangkan untuk parameter σ^2 plot fungsi densitasnya memiliki bentuk menyerupai plot fungsi densitas distribusi Inverse Gamma.

3.2 Data Penundaan Penerbangan di US tahun 2008

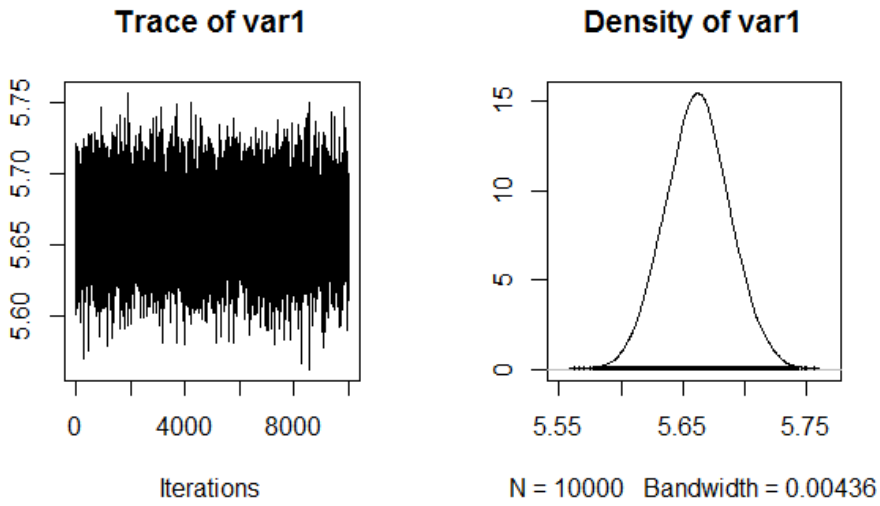
Data real yang digunakan adalah data penundaan penerbangan yang diambil dari http://stat-computing.org/dataexpo/2009/the-data.html periode tahun 2008 dan disediakan oleh U.S. Department of Transportation. Beberapa faktor yang menyebabkan *delay* secara garis besar yaitu manajemen airline, faktor teknis operasional, faktor cuaca, dan faktor lain seperti kerusuhan/demontrasi di wilayah bandara. Pada kasus ini, diambil variabel *Late Arrival Delay*, *Air Time*, *Departure Delay*, *Carrier Delay*, *Weather Delay*, *NAS Delay*, dan *Security Delay* yang dianggap sebagai faktor penyebab keterlambatan pesawat yang diukur dengan variabel *ArrDelay*, dengan total observasi sebesar 1.936.758 data. Sebelum dilakukan analisis menggunakan regresi linier Bayesian, terlebih dahulu dilakukan preprocessing data yaitu menghilangkan missing value. Cara menghilangkan missing value dilakukan dengan bantuan software R. Sehingga total observasi setelah menghilangkan missing value sebesar 689.270 data.

Ilustrasi penggunaan paket program *R* untuk data penundaan penerbangan ini secara garis besar sama dengan data simulasi yang sudah dijelaskan pada sub-bab D. Fungsi untuk model regresi linear Bayesian dengan `bayes.regress()`. Terlebih dahulu ditentukan distribusi prior untuk $\boldsymbol{\beta}$ dan σ^2, dalam hal ini dipilih distribusi multivariate normal dengan vektor mean diketahui dan matriks kovariansi tidak diketahui untuk distribusi prior $\boldsymbol{\beta}$ dan distribusi Invers Gamma dengan parameter shape dan scale diketahui untuk distribusi prior σ^2. Nilai yang dihasilkan sim.beta.sigmasq.out merupakan matriks sampel posterior MCMC untuk $\boldsymbol{\beta}$ yang berdimensi (Tsamp.out, k+1), dan vektor sampel posterior MCMC untuk σ^2 yang berdimensi (Tsamp.out). Tsamp.out ditetapkan sebesar 11.000. Analisis sampel posterior MCMC dapat dilihat dari plot densitas untuk distribusi posterior marginal pada masing-masing β, dan ringkasan statistik untuk distribusi posterior marginal β dapat dilihat secara ringkas pada Tabel 2.

Tabel 2. Prediksi nilai posterior mean dan kuantil posterior untuk parameter model yang tidak diketahui pada data penundaan penerbangan di US tahun 2008

Parameter	Posterior Mean	Posterior 2,5%	Posterior 25%	Posterior 50%	Posterior 75%	Posterior 97,5%
β_0	5,6618665	5,611	5,644	5,662	5,679	5,714
β_1	0,01756	0,01722	0,01744	0,01756	0,01768	0,0179
β_2	0,7927	0,7922	0,7926	0,7927	0,7929	0,7933
β_3	0,1544	0,1537	0,1542	0,1544	0,1547	0,1552
β_4	0,2174	0,2161	0,2169	0,2174	0,2179	0,2188
β_5	0,4425	0,4417	0,4422	0,4425	0,4428	0,4433
β_6	0,1033	0,09004	0,09878	0,10328	0,10787	0,11628
σ^2	234,8	234,3	234,6	234,8	235	235,3

Dari tabel 2. menunjukkan bahwa penduga dari masing-masing parameter dapat dilihat pada kolom posterior mean. Selanjutnya kuantil posterior 2,5% dan 97,5% menunjukkan batas bawah dan batas atas dari interval kredibel (interval kepercayaan Bayesian) untuk masing-masing parameter.



Gambar 2. Iterasi Gibbs Sampler dan Plot Fungsi Densitas untuk β_0 pada data penundaan penerbangan di US tahun 2008

Gambar 2. menunjukkan iterasi Gibbs sampler pada β_0, selanjutnya plot fungsi densitas untuk parameter β_0 memiliki bentuk menyerupai plot fungsi densitas distribusi Normal, demikian seterusnya sampai β_6. Sedangkan untuk parameter σ^2 plot fungsi densitasnya memiliki bentuk menyerupai plot fungsi densitas distribusi Inverse Gamma.

IV. KESIMPULAN

Perkembangan terbaru dalam ilmu data dan penelitian, menghasilkan dataset besar yang terlalu besar untuk dianalisis secara keseluruhan karena keterbatasan memori komputer atau kapasitas penyimpanan. Untuk mengatasi hal tersebut diperkenalkan paket program dari R yaitu `BayesSummaryStatLM` untuk model regresi linier Bayesian dengan implementasi Markov chain Monte dan paket program dari R yaitu `ff` digunakan untuk membaca data pada dataset besar sekaligus menghitung ringkasan statistik. Pada data simulasi dan data real yang telah dianalisis diperoleh bahwa penduga dari masing-masing parameter dapat dilihat pada posterior mean. Selanjutnya kuantil posterior 2,5% dan 97,5% menunjukkan batas bawah dan batas atas dari interval kredibel (interval kepercayaan Bayesian) untuk masing-masing parameter. Pada data simulasi, penduga untuk masing-masing parameter yang dihasilkan mempunyai nilai yang mendekati nilai parameter yang ditentukan (True Value). Hal ini juga ditunjukkan oleh sempitnya interval kredibel untuk masing-masing parameter.

REFERENSI

[1] R Core Team, "R: A Languange and Environment for Statistical Computing," The R Project for Statistical Computing website, 2014. [Online]. Available: http://www.R-project.org/. [Accessed 11 March 2017].

[2] Adler D, Glaser C, Nenadic O, Oehlschlagel J, Z, "ff: memory-efficient storage of large data on disk and fast access functions," R package version 2.2-13, 2013. [Online]. Available: http://CRAN.R-project.org/package=ff. [Accessed 10 March 2017].

[3] Ordonez C, Garcia-Alvarado C, Baladandayuthapan, "Bayesian variable selection in linear regression in one pass for large data sets," *ACM Transactions on Knowledge Discovery from Data,* vol. 9(1), no. 3, p. doi: 10.1145/26296178, 2014.

[4] Ghosh J, Reiter JP, "Secure Bayesian model averaging for horizontally partitioned data," *Statistics and Computing,* vol. 23, pp. 311-322, 2013.

[5] Carlin BP, Louis TA, Bayesian Methods for Data Analysis. 3rd ed, Boca Raton: FL: Chapman and Hall/CRC Press, 2009.

[6] Gelman A, Carlin JB, Stern HS, Dunson DB, Vehta, Bayesian Data Analysis 3rd ed, Boca Raton: FL: Chapman and Hall/CRC Press, 2013.

[7] Alexey M, Evgeny S, Erin M C, "BayesSummaryStatLM: An R package for Bayesian Linear Models for Big data and Data Science".

[8] United States Department of Transportation, "Bureau of Transportation Statistics," [Online]. Available: http://stat-computing.org/dataexpo/2009/the-data.html. [Accessed 25 September 2017].

[9] Evgeny Savel'ev, Alexey Miroshnikov, Erin Conlon, "MCMC Sampling of Bayesian Linear Models via Summary Statistics", 2015.

[10] Robert CP, Casella G, Monte Carlo Statistical Methods, 2nd ed, New York, NY: Springer, 2004.

[11] Lindley DV, Smith AFM, "Bayes estimates for the linear model", J R Stat Soc B, 1972, vol 34, pp. 1-41.