JURNAL GAUSSIAN, Volume 2, Nomor 1, Tahun 2013, Halaman 79-88
Online di: http://ejournal-s1.undip.ac.id/index.php/gaussian

ESTIMASI PARAMETER REGRESI LOGISTIK MULTINOMIAL DENGAN METODE BAYES

Wayaning Apsari[1], Hasbi Yasin[2], Sugito[3]
[1]Mahasiswa Jurusan Statistika FSM Universitas Diponegoro
[2,3]Staf Pengajar Jurusan Statistika FSM UNDIP

ABSTRAK

Regresi logistik multinomial merupakan regresi logistik dimana variabel dependennya bersifat *polychotomous* yaitu nilai variabel dependennya lebih dari dua kategori. Pada umumnya estimasi parameter regresi logistik multinomial menggunakan metode klasik yang hanya didasarkan pada informasi saat ini yang diperoleh dari sampel tanpa memperhitungkan informasi awal dari parameter regresi logistik. Jika dimiliki informasi awal tentang parameter yaitu distribusi prior, maka estimasi parameter dapat menggunakan metode Bayes. Metode Bayes menggabungkan informasi pada sampel dengan informasi distribusi prior, dan hasilnya dinyatakan dengan distribusi posterior. Jika distribusi posteriornya tidak dapat diturunkan secara analitis maka didekati dengan menggunakan algoritma *Markov Chain Monte Carlo* (MCMC) terutama algoritma *Metropolis-Hastings.* Algoritma ini menggunakan mekanisme penerimaan dan penolakan untuk membangkitkan barisan sampel random.

Kata kunci: Regresi Logistik Multinomial, Metode Bayes, algoritma *Markov Chain Monte Carlo* (MCMC), algoritma *Metropolis-Hastings*.

ABSTRACT

Multinomial logistic regression is a logistic regression where the dependent variable is *polychotomous* is dependent variable value of more than two categories. Multinomial logistic regression parameter estimation usually use classical method that is based only on current information obtained from the sample without taking into account the initial information of logistic regression parameters. If have early information about parameter is prior distribution, the parameter estimation can use Bayes method. Bayesian methods combine information on the sample with prior distribution of information, and the results are expressed in the posterior distribution. If posterior distribution can not be derived analytically so approximated using *Markov Chain Monte Carlo* (MCMC) algorithm especially *Metropolis-Hastings* algorithm. This algorithm uses acceptance and rejection mechanism to generate a sequence of random samples.

Keyword*:* Multinomial Logistic Regression, Bayes Method, Markov Chain Monte Carlo algorithm (MCMC), Metropolis-Hastings algorithm.

1. PENDAHULUAN

Regresi logistik multinominal merupakan regresi logistik dimana variabel dependennya mempunyai skala yang bersifat *polychotomous* atau multinominal yang terdiri lebih dari dua kategori. Pendugaan koefisien parameter model regresi logistik multinomial pada umumnya menggunakan metode Maksimum Likelihood dengan menggunakan pendekatan distribusi. Pada umumnya metode klasik ini hanya berkutat pada informasi saat ini yang diperoleh dari sampel tanpa memperhitungkan informasi awal dan hanya mendasarkan inferensinya pada sampel. Sehingga jika distribusi populasi tidak diketahui metode Maksimum Likelihood tidak dapat digunakan.

Inferensi akan lebih bagus jika data yang digunakan adalah data gabungan antara data sampel saat ini dengan data penelitian sebelumnya (data prior). Metode inferensi dengan menggunakan data sampel dan data prior disebut dengan metode Bayes [1]. Distribusi prior adalah distribusi subyektif berdasarkan pada keyakinan seseorang dan dirumuskan sebelum data sampel diambil[2]. Distribusi sampel yang digabung dengan distribusi prior akan menghasilkan distribusi baru yaitu distribusi posterior.

Kepadatan posterior untuk parameter regresi pada model multinomial tidak dapat diturunkan secara analitis. Sebaliknya, teknik numerik diperlukan untuk meringkas distribusi peluang ini. Karena penyelesaian untuk estimasi marginal posterior setiap parameter dari persamaan itu akan rumit, sehingga akan didekati dengan algoritma *Markov Chain Monte Carlo* terutama algoritma *Metropolis-Hastings.*

2. TINJAUAN PUSTAKA

2.1 Regresi Logistik Multinomial

Regresi logistik multinomial merupakan regresi logistik dengan variabel dependen (Y) mempunyai skala yang bersifat *polychotomus* atau multinomial yaitu skala dengan kategori lebih dari dua[3].
Misal **X** variabel independen yang berukuran (p+1) dan variabel dependen **Y** (j kategori) mempunyai kategori j = 0, 1, 2 dengan probabilitas respon π_0, π_1, π_2 dan $\sum_{j=0}^{2} \pi_j = 1$

Probabilitas bersyarat $P(y = j| x) = \pi_j(x)$, j =0, 1, 2
Jadi probabilitas bersyarat $P(y = j|x) = \pi_j(x)$, j = 0,1,2 dapat ditulis:

$$\pi_0(x) = \frac{1}{1 + e^{g_1(x)} + e^{g_2(x)}}$$

$$\pi_1(x) = \frac{e^{g_1(x)}}{1 + e^{g_1(x)} + e^{g_2(x)}}$$

$$\pi_2(x) = \frac{e^{g_2(x)}}{1 + e^{g_1(x)} + e^{g_2(x)}}$$

Dengan fungsi logit sebagai berikut:

$$g_1(x) = \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2$$

$$g_2(x) = \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2$$

2.2 Teorema Bayesian

Misal peristiwa-peristiwa $A_1, A_2, \ldots, A_k$ membentuk partisi di ruang sampel S sedemikian hingga $P(A_i) > 0$; i=1, 2,…,k dan misalkan B sebarang peristiwa sedemikian hingga $P(B) > 0$. Maka untuk i=1, 2, …,k

$$P(A_j|B) = \frac{P(A_j)\,P(B|A_j)}{\sum_{i=1}^{k} P(A_i)P(B|A_i)}$$

Teorema bayes memberikan aturan sederhana untuk menghitung probabilitas bersyarat peristiwa A_i diberikan B terjadi, jika masing-masing probabilitas tak bersyarat A_i dan probabilitas bersyarat B diberikan A_i terjadi diketahui[4].

2.2.1 Distribusi Prior

Distrribusi prior dikelompokkan menjadi dua berdasarkan bentuk fungsi likelihood, yaitu[5]:

1. Berkaitan dengan bentuk distribusi hasil identifikasi pola datanya
   a. Distribusi prior konjugat (*conjugate*), mengacu pada acuan analisis model terutama dalam pembentukan fungsi likelihoodnya sehingga dalam penentuan prior konjugat selalu dipikirkan mengenai penentuan pola distribusi prior yang mempunyai bentuk konjugat dengan fungsi densitas peluang pembangun likelihoodnya.
   b. Distribusi prior tidak konjugat (*non-conjugate*), pemberian prior pada model tidak mempertimbangkan pola pembentuk fungsi likelihoodnya

2. Berkaitan dengan penentuan parameter pada pola distribusi prior
   a. Distribusi prior informatif, mengacu pada pemberian parameter dari distribusi prior yang telah dipilih baik distribusi prior konjugat atau tidak, pemberian nilai parameter pada distribusi prior ini didasarkan pada informasi yang diperoleh
   b. Distribusi prior non informatif, pemilihannya tidak didasarkan pada data yang ada atau distribusi prior yang tidak mengandung informasi tentang parameter θ.
      Apabila pengetahuan tentang priornya sangat lemah, maka bisa digunakan prior berdistribusi normal dengan mean nol dan varian besar. Efek dari penggunaan prior dengan mean nol adalah estimasi parameternya dihaluskan menuju nol. Pemulusan ini dilakukan oleh varian, sehingga pemulusan tersebut bisa dilakukan dengan meningkatkan varian[6].

2.2.2 Distribusi Posterior

Distribusi posterior adalah fungsi densitas bersyarat θ jika diketahui nilai observasi x dan dapat ditulis sebagai berikut[4]:

$$f(\theta \mid x) = \frac{f(\theta, x)}{f(x)}$$

Fungsi kepadatan bersama dan marginal yang diperlukan dapat ditulis dalam bentuk distribusi prior dan fungsi likelihood,

$$f(\theta, x) = f(x \mid \theta) f(\theta)$$

$$f(x) = \int_{-\infty}^{\infty} f(\theta, x) d\theta = \int_{-\infty}^{\infty} f(\theta) f(x \mid \theta) d\theta$$

Sehingga fungsi densitas posterior untuk variabel random kontinu sebagai berikut,

$$f(\theta \mid x) = \frac{f(\theta) f(x \mid \theta)}{\int_{-\infty}^{\infty} f(\theta) f(x \mid \theta) d\theta}$$

2.3 Algoritma Metropolis-Hastings

Persamaan posterior yang mempunyai bentuk analitik yang sulit, untuk mengetahui nilai estimasi parameter dari bentuk tersebut akan digunakan simulasi *Random-walk Metropolis-Hastings*. Sebelum memulai iterasi, terlebih dahulu ditentukan distribusi proposal yang akan digunakan[7].

Langkah-langkah dari simulasi Random-walk Metropolis-Hastings akan berjalan sebagai berikut:

1. Menentukan nilai awal
2. Menentukan banyak iterasi t=1,…,T
   a. Mengatur $\beta = \beta^{t-1}$
   b. Membangkitkan nilai baru β' dari dari distribusi proposal $N(\beta, \bar{s}_{\beta}^2)$.
   c. Menghitung $\log \alpha = \min(0, A)$, dengan A diberikan oleh
      $$A = log \frac{L(y|\beta')g(\beta')}{L(y|\beta)g(\beta)}$$
   d. Membangkitkan sampel random $u \sim U(0,1)$
   e. Memperbaharui $\beta^t = \beta'$ dengan peluang penerimaan α dan $\beta^t = \beta^{t-1}$ dengan peluang 1-α. Jika $u \leq \alpha$ maka β' diterima sebagai anggota sampel dan jika $u > \alpha$ maka nilai sebelumnya (β) yang diterima sebagai anggota sampel.

3. PEMBAHASAN

3.1 Fungsi Likelihood

Pada model regresi logistik multinomial, Y_i terdiri lebih dari dua kategori maka model regresi logistik multinomial didasarkan pada distribusi multinomial $y_i \sim M(n, \pi)$

Fungsi densitas peluang untuk regresi logistik multinomial dengan tiga kategoti adalah

$$f(y \mid \beta) = \pi_0(x_i)^{y_{0i}} . \pi_1(x_i)^{y_{1i}} . \pi_2(x_i)^{y_{2i}}$$

Fungsi likelihood untuk data $Y = \{y_1, y_2, .., y_n\}$ adalah sebagai berikut

$$L(\beta \mid y) = \exp\left\{ g_1(x)\sum_{i=1}^{n} y_{1i} + g_2(x)\sum_{i=1}^{n} y_{2i} - \sum_{i=1}^{n} \log\left(1 + e^{g_1(x)} + e^{g_2(x)}\right)\right\}$$

dengan

$$g_1(x) = \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2$$
$$g_2(x) = \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2$$

3.2 Distribusi Prior

Distribusi prior Normal untuk model regresi logistik multinomial adalah

$$g(\beta_P) = \frac{1}{\sqrt{2\pi\sigma_P^2}} \exp\left\{ -\frac{1}{2}\left(\frac{\beta_p - \mu_p}{\sigma^p} \right)^2 \right\}$$

3.3 Distribusi Posterior

$$g(\beta \mid y) = \exp\left\{ g_1(x)\sum_{i=1}^{n} y_{1i} + g_2(x)\sum_{i=1}^{n} y_{2i} - \sum_{i=1}^{n} \log\left(1 + e^{g_1(x)} + e^{g_2(x)}\right) + \left\{ -\frac{1}{2}\left(\frac{\beta_p - \mu_{\beta_p}}{\sigma_{\beta_p}} \right)^2 \right\}\right\}$$

Distribusi posterior yang digunakan untuk mengestimasi parameter regresi pada model multinomial mempunyai bentuk analitik yang sulit. Untuk itu dilakukan simulasi dari distribusi posterior yang terbentuk. Metode simulasi yang digunakan adalah algoritma *Markov Chain Monte Carlo* khususnya *Metropolis Hastings*.

Untuk mengimplementasikan algoritma Metropolis-Hastings perlu ditentukan distribusi proposal yang tepat. Jika distribusi proposal simetris maka pengambilan sampel dengan *Random-walk Metropolis Hastings sampling*. Distribusi proposal yang digunakan untuk regresi logistik multinomial untuk tiga kategori dan dua variabel independen menggunakan *Independent Normal proposal* adalah

$$\beta^{'} \sim N_6(\beta, diag\left(\bar{s}^2_{\beta_{10}}, \bar{s}^2_{\beta_{11}}, \bar{s}^2_{\beta_{12}}, \bar{s}^2_{\beta_{20}}, \bar{s}^2_{\beta_{21}}, \bar{s}^2_{\beta_{22}}\right)$$

3.5 Contoh Aplikasi

Data diambil dari buku[8], halaman 388-389. Sebanyak 63 sampel Aligator di Danau George, dimana setiap aligator mempunyai pilihan makanan utama yang berbeda, yaitu ikan, siput atau cacing, dan lainnya (katak, kura-kura, ular, burung, ular, reptil, dan mamalia). Sebagai variabel independen adalah panjang dan jenis kelamin aligator. Panjang aligator diklasifikasikan secara biner yaitu jika panjang aligator ≤ 1.83 meter maka dikategorikan aligator muda, jika panjang aligator > 1.83 meter maka dikategorikan aligator dewasa sedangkan jenis kelamin dikategorikan menjadi jantan dan betina

3.5.1 Distribusi Prior

$$g(\beta_{10},\beta_{11},\beta_{12},\beta_{20},\beta_{21},\beta_{22}) = \frac{1}{\left(\sqrt{20000\pi}\right)^6}\exp\left\{-\frac{\beta_{10}^2}{20000}-\frac{\beta_{11}^2}{20000}-\frac{\beta_{12}^2}{20000}-\frac{\beta_{20}^2}{20000}-\frac{\beta_{21}^2}{20000}-\frac{\beta_{22}^2}{20000}\right\}.$$

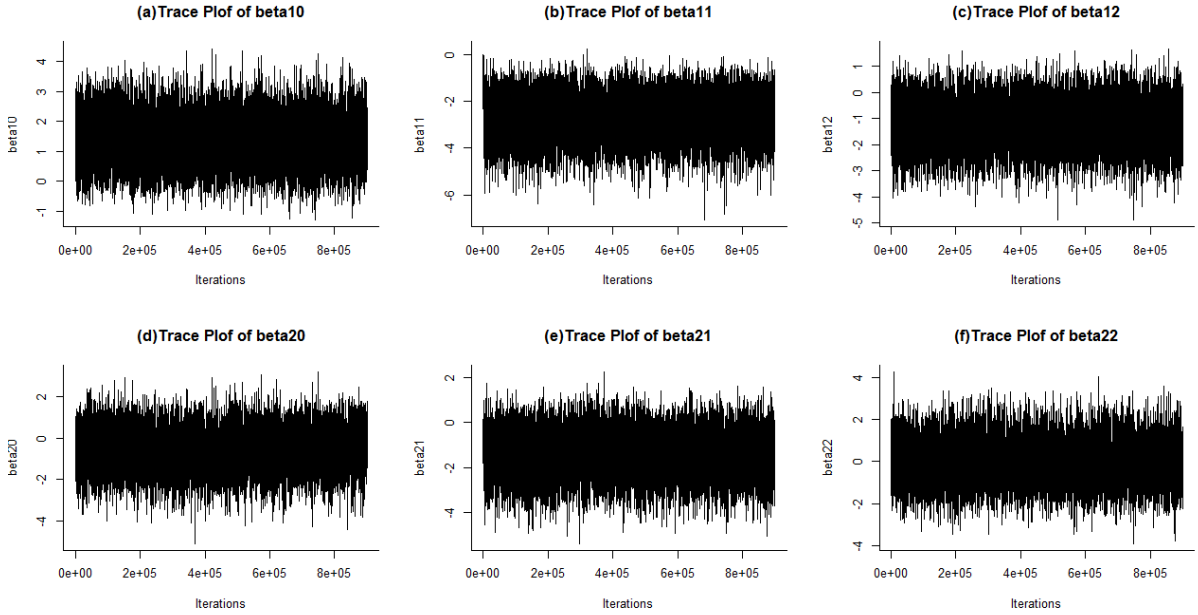
3.5.2 Distribusi Posterior

$$g(\beta_{10},\beta_{11},\beta_{12},\beta_{20},\beta_{21},\beta_{22}\mid y) \propto \exp\left\{\begin{array}{l}\left[g_1(x)\sum_{i=1}^{n}y_{1i}+g_2(x)\sum_{i=1}^{n}y_{2i}-\sum_{i=1}^{n}\log\left(1+e^{g_1(x)}+e^{g_2(x)}\right)\right] \\ +\left[-\frac{\beta_{10}^2}{20000}-\frac{\beta_{11}^2}{20000}-\frac{\beta_{12}^2}{20000}-\frac{\beta_{20}^2}{20000}-\frac{\beta_{21}^2}{20000}-\frac{\beta_{22}^2}{20000}\right]\end{array}\right\}$$

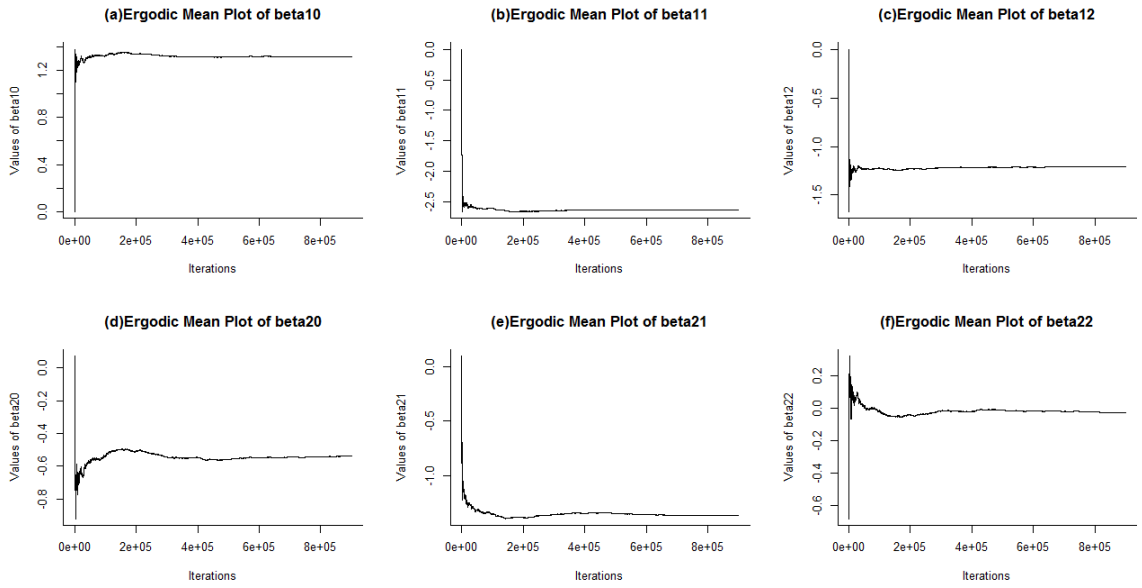
Distribusi posterior yang digunakan untuk mengestimasi parameter regresi logistik multinomial mempunyai bentuk analitik yang sulit. Untuk itu dilakukan simulasi dari distribusi posterior yang terbentuk. Jalannya simulasi tersebut membutuhkan nilai prior, nilai awal, dan distribusi proposal.

1. Prior
   Untuk mengatasi sedikitnya informasi, maka digunakan prior berdistribusi normal (0, 100^2)
2. Nilai awal
   Nilai awal yang digunakan dalam proses simulasi semua paramter adalah 0
3. Distribusi Proposal
   Distribusi proposal yang digunakan adalah independent normal proposal dengan nilai $\bar{s}_{\beta_p} = 1$

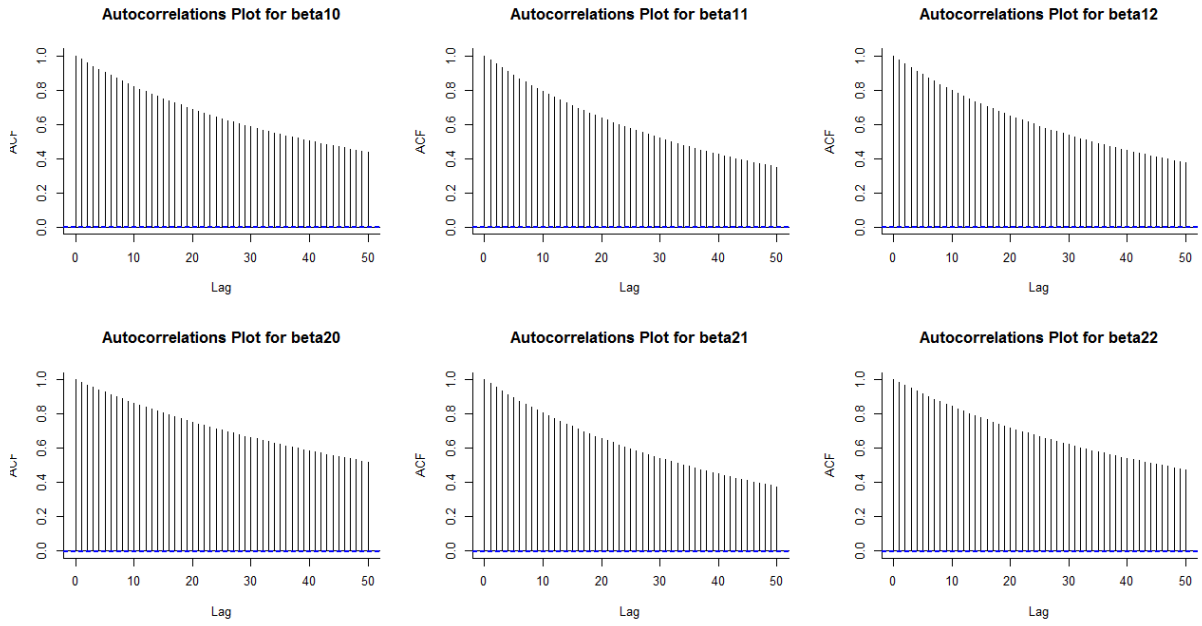
Langkah selanjutnya adalah menjalankan simulasi Random-walk Metropolis Hastings dengan iterasi awal sebanyak 50.000 iterasi tetapi memberikan hasil yang belum konvergen. Untuk mengatasi hal tersebut yaitu dengan menambah iterasi dan iterasi meningkat sampai 900.000 untuk memastikan konvergensi.



Gambar 1 Trace plot sebanyak 900.000 iterasi

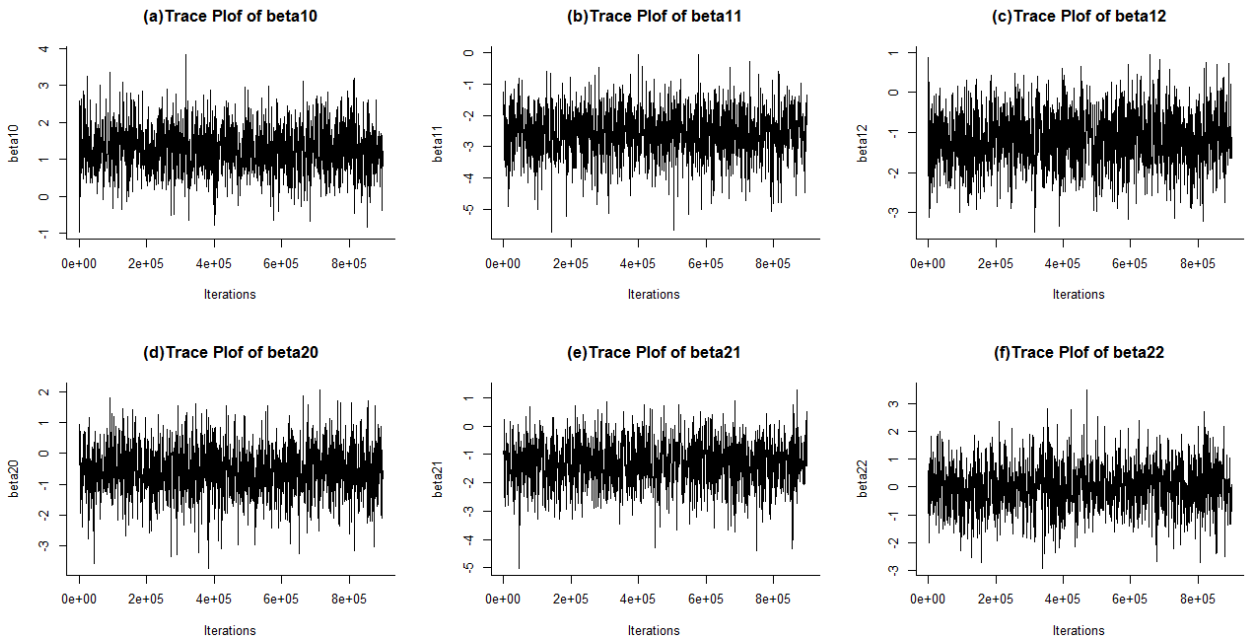


Gambar 2 Ergodic mean plot sebanyak 900.000 iterasi

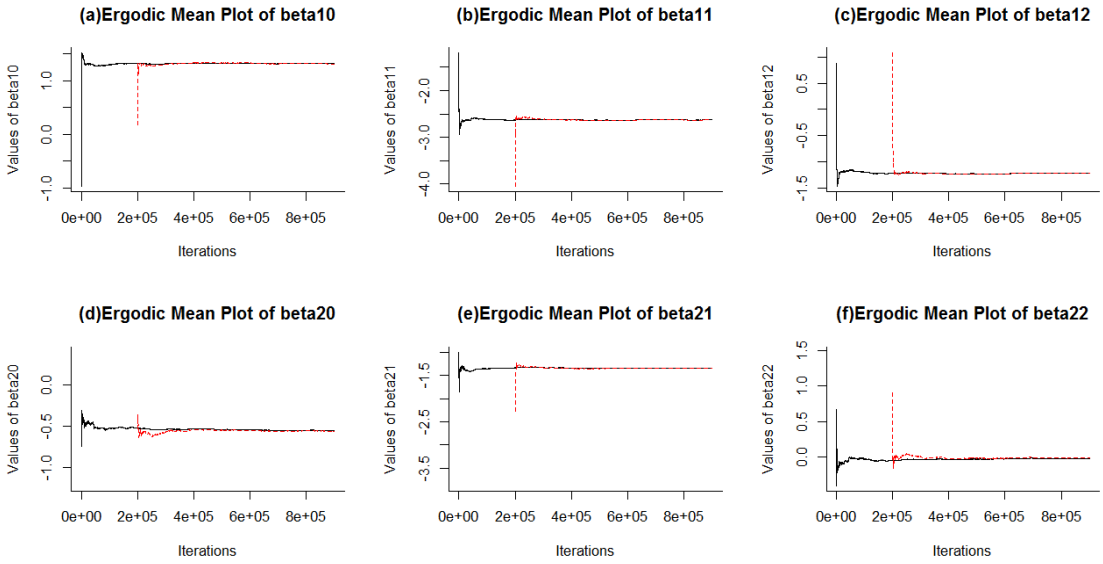


Gambar 3 Plot autokorelasi sebanyak 900.000 iterasi

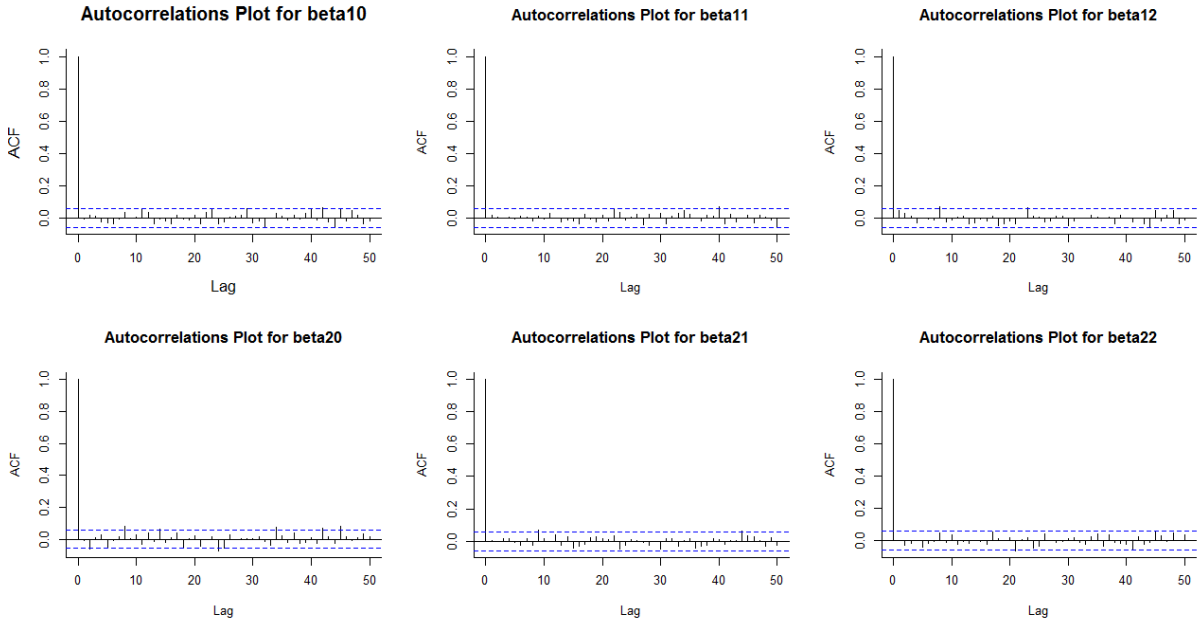
Setelah kondisi konvergen terpenuhi, langkah selanjutnya adalah mencari nilai estimasi parameter beta. Untuk menghindari nilai awal, maka iterasi ini akan dimulai pada iterasi ke 200.001 dimana kondisi mulai dari iterasi ini sudah menunjukkan konvergen.



Gambar 4 Trace plot dengan burnin 200.000 dan thin 600



Gambar 5 Ergodic mean plot dengan burnin 200.000 dan thin 600



Gambar 6 Plot autokorelasi dengan burnin 200.000 dan thin 600

Gambar 4,5 dan 6 merupakan trace plot, ergodic mean plot, dan plot autokorelasi sebanyak 900.000 iterasi dengan burnin 200.000 dan thinning interval 600. Setelah iterasi 0-200.000 dihilangkan, maka didapatkan nilai estimasi parameter regresi logistik multinomial yang baru.

3.5.3 Pembentukan Model

Pengujian hipotesis terhadap parameter regresi dilakukan dengan pendekatan interval konfidensi 95% dari masing-masing parameter. Hal ini dikarenakan distribusi posterior tidak diketahui dengan pasti. Interval konfidensi 95% dihitung dengan batas bawah yaitu kuantil ke 2,5% dan batas atasnya adalah kuantil ke 97,5%. Parameter dinyatakan signifikan jika interval konfidensi 95% parameter tidak memuat nilai nol[7].

Tabel 1 Nilai Estimasi Parameter

Variabel	Parameter	Mean	2,5% Kuantil	97,5% Kuantil	Signifikan	Kesimpulan
Konstanta	β_{10}	1.3108	0.07987475	2.70461989	-	-
Panjang	β_{11}	-2.6266	-4.258373	-1.203736	ya	Berpengaruh
Jenis Kelamin	β_{12}	-1.2093	-2.6716109	0.1781262	Tidak	Tidak Berpengaruh
Konstanta	β_{20}	-0.55212	-2.305870	1.092261	-	-
Panjang	β_{21}	-1.3372	-2.987810	0.190822	Tidak	Tidak Berpengaruh
Jenis Kelamin	β_{22}	-0.02206	-1.677998	1.730372	Tidak	Tidak berpengaruh

Dari tabel di atas diketahui variabel yang berpengaruh hanya panjang dan variabel jenis kelamin tidak berpengaruh, sehingga yang dimasukkan ke dalam model hanya variabel panjang. Sehingga didapat model sebagai berikut

Fungsi Logit:

$$g_1(x) = 1.3108 - 2.6266P$$

$$g_2(x) = -0.55212 - 1.3372P$$

Nilai Probabilitas:

Untuk pilihan makanan ikan

$$\pi_1(x) = \frac{e^{1.3108-2.6266P}}{1+e^{1.3108-2.6266P}+e^{-0.55212-1.3372P}}$$

Untuk pilihan makanan siput atau cacing

$$\pi_2(x) = \frac{e^{-0.55212-1.3372P}}{1+e^{1.3108-2.6266P}+e^{-0.55212-1.3372P}}$$

Untuk pilihan makanan lainnya

$$\pi_0(x) = \frac{1}{1+e^{1.3108-2.6266P}+e^{-0.55212-1.3372P}}$$

Contoh perhitungan:

Seekor aligator mempunyai panjang 1.30 meter akan dicari peluangnya memilih makanan utama ikan, siput atau cacing, dan makanan lain.

Panjang aligator =1.30 meter dikoding 0

a. Probabilitas memilih makanan ikan

$$\pi_1(x) = \frac{e^{1.3108-2.6266P}}{1+e^{1.3108-2.6266P}+e^{-0.55212-1.3372P}}$$

$$= \frac{e^{1.3108-2.6266(0)}}{1+e^{1.3108-2.6266(0)}+e^{-0.55212-1.3372(0)}}$$

$$= 0.7018$$

b. Probabilitas memilih makanan siput atau cacing

$$\pi_2(x) = \frac{e^{-0.55212-1.3372P}}{1+e^{1.3108-2.6266P}+e^{-0.55212-1.3372P}}$$

$$= \frac{e^{-0.55212-1.3372(0)}}{1+e^{1.3108-2.6266(0)}+e^{-0.55212-1.3372(0)}}$$

$$= 0.1090$$

c. Probabilitas memilih makanan lain

$$\pi_0(x) = \frac{1}{1+e^{1.3108-2.6266P}+e^{-0.55212-1.3372P}}$$

$$= \frac{1}{1+e^{1.3108-2.6266(0)}+e^{-0.55212-1.3372(0)}}$$

$$= 0.1892$$

Jadi, seeekor aligator yang mempunyai panjang 1.30 meter mempunyai probabilitas memilih makanan ikan sebesar 0.7018, probabilitas memilih makanan siput atau cacing sebesar 0.1090 dan probabilitas memilih makanan lain sebesar 0.1892. Ini berarti, seekor aligator yang mempunyai panjang ≤ 1.83 meter cenderung memilih makanan ikan.

4. KESIMPULAN

1. Jika diketahui pengetahuan awal tentang parameter regresi logistik multinomial yang dinyatakan dengan distribusi prior, maka estimasi parameter dapat dilakukan dengan menggunakan metode Bayes.
2. Jika distribusi posterior dari parameter regresi logistik multinomial sulit diselesaikan secara analitik, maka digunakan algoritma *Markov Chain Monte Charlo* terutama *Metropolis Hastings*. Algoritma ini menggunakan mekanisme penerimaan dan penolakan untuk membangkitkan barisan sampel random.

DAFTAR PUSTAKA

1. Bolstad, W.M. 2007. *Introduction to Bayesian Statistics Second Edition.* A John Wiley & Sons. Inc: America.
2. Walpole, R. E. dan Myers, R. H. 1986. *Ilmu Peluang dan Statistika untuk Insinyur dan Ilmuwan*. Terbitan kedua. ITB: Bandung.
3. Hosmer, D.W. and Lemeslow. 2000. *Applied Logistic Regression Second Edition*. John Wiley & Sons, Inc: New York.
4. Soejati, Z dan Soebanar. 1998. Inferensi Bayesian. Karunia Universitas Terbuka; Jakarta.
5. Box, G.E.P and Tiao, G.C. 1973. *Bayesian Inference In Statistical Analysis*. Addision-Wesley Publishing Company, Inc: Philippines.
6. Ntzoufras, I. 2009. *Bayesian Modelling Using WinBUGS.* John Wiley & Sons, Inc: Ney Jersey.
7. Galindo-Garre, F. and Vermunt, J. K. 2004. *Bayesian Posterior Estimation of Logit Parameters With Small Samples, Artikel*. Sage Publication: Netherlands.
8. Agresti, A. 1996. *An Introduction to Categorial Data Analysis*.New York: John Wiley & Son's.