ISBN: 978-602-19356-2-0

(16)

prosiding

Seminar Nasional Statistika, Matematika dan Aplikasinya 2014

SNSMA 2014

Dipublikasikan oleh:
Fakultas Matematika & Ilmu Pengetahuan Alam
Universitas Islam Bandung

Prosiding Seminar Nasional Statistika, Matematika dan Aplikasinya
Universitas Islam Bandung, 2014, Halaman : 19-26

Analisis Bayesian pada Regresi Binomial dengan Kesalahan Klasifikasi

Retno Budiarti
Departemen Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Institut Pertanian Bogor
E-mail : retno.budiarti@gmail.com

ABSTRAK

Dalam tulisan ini, difokuskan pada analisis Bayesian dari regresi binomial dengan kesalahan klasifikasi pada peubah respon dan *error-free* pada peubah-peubah penjelas (*covariates*). Pendekatan data augmented memberikan kemudahan bentuk sebaran prior dan memberikan kesimpulan untuk parameter-parameter yang menggambarkan hubungan antara peubah-peubah penjelas dengan peubah respon dan untuk kesalahan pada peluang. Selanjutnya, pendekatan ini berlaku untuk setiap model linier terampat (*generalized linear model*). Pemilihan *link function* cukup mempengaruhi pendugaan parameter. Secara umum, kesalahan klasifikasi menyebabkan pendugaan parameter menjadi tidak valid.

Kata-kata kunci: generalized linear model, regresi binomial, analisis Bayesian, kesalahan klasifikasi.

1. PENDAHULUAN

Model linier terampat terdiri atas tiga komponen yaitu komponen acak, komponen sistematis, dan fungsi hubung (*link function*) (McCullagh dan Nelder, 1987). Pada tulisan ini, bertindak sebagai komponen acak adalah peubah respon yang memiliki hanya dua kemungkinan nilai yaitu 0 (gagal) dan 1 (sukses). Oleh karena itu, diusulkan untuk menggunakan regresi binomial.

Klasifikasi muncul secara natural dalam banyak situasi. Ketika informasi dikumpulkan, biasanya data tidak bebas dari kesalahan (*misclassification*). Kesalahan ini dapat terjadi dikarenakan beberapa sebab, seperti dalam kegiatan survei konsumen : konsumen tidak mengerti pertanyaan dalam kuesioner, atau terjadi kesalahan pencatatan secara intensif.

Dalam tulisan ini, diusulkan menggunakan analisis Bayesian pada regresi binomial dengan misklasifikasi pada peubah respon dan bebas kesalahan (*error-free*) pada peubah-peubah penjelas. Artinya, adanya kesalahan klasifikasi pada peubah respon tidak ada hubungannya dengan peubah penjelas. Kesalahan klasifikasi memberi dampak pada pengambilan kesimpulan, bahkan kesalahan klasifikasi pada sebagian kecil data sekalipun (Paulino *et al*, 2005).

Permasalahannya adalah seberapa besar kesalahan pada pengambilan kesimpulan (hasil pendugaan parameter) yang ditimbulkan kesalahan klasifikasi pada peubah respon. Adakah pemilihan *link function* pada kasus regresi binomial dengan kesalahan klasifikasi juga berpengaruh pada pendugaan parameter. Selain itu, apakah banyaknya ukuran sampel berpengaruh juga pada pendugaan parameter. Berdasarkan permasalahan tersebut, tujuan dari tulisan ini adalah mencari tahu pemilihan *link function* dan ukuran sampel pada regresi binomial dalam menentukan pengambilan keputusan (hasil pendugaan parameter), dengan menggunakan studi simulasi.

2. TINJAUAN PUSTAKA

2.1. Model Regresi Binomial dengan Kesalahan Klasifikasi

Misalkan data regresi $(n_k, N_k, x_k), k = 1, \ldots, N$, dimana n_k merupakan banyaknya kejadian sukses dari sebaran binomial saling bebas, Binomial (N_k, ϕ_k), x_k diketahui sebagai vektor peubah penjelas

berukuran $p \times 1$. Kesalahan klasifikasi peubah respon diakomodasi dengan membagi proses pengumpulan data ke dalam dua tahap : *an unobserved sampling stage* berkaitan dengan *true response* y^T diikuti oleh *a reporting stage* dimana pengamatan dan kemungkinan kesalahan y^0 dicatat.

Jika $y = 1$ diasosiasikan sebagai "sukses" dan $\theta_{ki} = P\left(y^T = i \middle| x_k\right), k = 1, \ldots, N, i = 0,1,$ dengan $\sum_i \theta_{ki} = 1$ dan $\lambda_{kij} = P\left(y^0 = j \middle| y^T = i, x_k\right), k = 1, \ldots, N, i, j = 0,1,$ dengan $\sum_j \lambda_{kij} = 1$, maka peluang sukses dari pengamatan respon individu dengan peubah penjelas x_k adalah $\phi_k = \sum_i \lambda_{ki1}\theta_{ki}$. Jadi model peluang untuk regresi data pengamatan $\mathbf{n} = (n_k, N_k)$ digambarkan sebagai perkalian likelihood binomial

$$L(\theta, \lambda | n) = \prod_{k=1}^{N} \binom{N_k}{n_k} \left(\sum_i \lambda_{ki1}\theta_{ki}\right)^{n_k} \left(\sum_i \lambda_{ki0}\theta_{ki}\right)^{N_k - n_k} \tag{1.1}$$

dengan θ dan λ adalah himpunan parameter berturut-turut dari θ_{ki} dan λ_{kij}. Metode standar yang menganalisis hubungan antara peubah respon dengan beberapa peubah penjelas adalah menggunakan model linier terampat (GLM). Dalam kasus ini, dapat diekspresikan nilai harapan proporsi sukses adalah

$$E\left(\frac{n_k}{N_k} \middle| \theta\right) = \theta_{k1} = \theta_1\left(x_k'\beta\right) = g\left(x_k'\beta\right),$$

dimana β adalah vektor koefisien regresi berukuran $p \times 1$, dan g(g) adalah fungsi distribusi kumulatif sebarang. Dalam kasus ini dipilih fungsi g(g) sebagai berikut

$$g(\mathbf{x}'\beta) = \begin{cases} \log\left(\exp\{\mathbf{x}'\beta\}\right) \\ \Phi(\mathbf{x}'\beta) \end{cases} \tag{1.2}$$

Selanjutnya, diasumsikan bahwa θ dan λ adalah saling bebas, dimotivasi oleh terjadinya kesalahan klasifikasi.

Bedrick *et al* (1996) mengusulkan metode untuk mengatasi masalah kesalahan klasifikasi ini menggunakan *conditional means prior* (CMP), yaitu

1) Pilih p vektor peubah penjelas $\tilde{x}_l; l = 1, \ldots, p;$
2) Menetapkan *prior* pada $\left\{\theta_1(\tilde{x}_1'\beta), \ldots, \theta_1(\tilde{x}_p'\beta)\right\};$
3) Dapatkan *the induced prior* pada β dengan menggunakan *the change-of-variables method.*

Menurut Paulino (2003) kehadiran kesalahan klasifikasi (*misclassification*) peubah respon, serta model peluang pada persamaan (1.1) menerangkan bahwa kesalahan klasifikasi menyebabkan distribusi posterior menjadi kompleks yang membuat tidak mungkin menyimpulkan secara langsung menggunakan metode analitik. Kita akan melihat pada subbab berikutnya bagaimana penggunaan data augmentation dapat mengurangi masalah ini dengan cara memisahkan parameter β dan λ dalam likelihoodnya.

2.2. Data Augmentation

Misalkan m_{kij} banyaknya pengamatan dengan $y^T = i$ dan $y^0 = j$ diantara pengamatan- pengamatan tersebut dengan peubah penjelas $\mathbf{x}_k$. Kita mempunyai $m_{k\cdot 1} = \sum_i m_{ki1} = n_k$ dan $m_{k\cdot 0} = N_k - n_k$, *the augmented data* $\mathbf{m} = (m_{kij})$ adalah sampel hipotetik dari distribusi multinomial $M\left\{N_k, (\lambda_{ki}\theta_{ki})\right\}$ dengan likelihood,

$$L(\beta, \lambda | m) \propto \prod_{k,i} \left\{\theta_i\left(x_k'\beta\right)\right\}^{m_{ki\cdot}} \prod_{k,i,j} \lambda_{kij}^{m_{kij}}.$$

Likelihood ini menunjukkan bahwa pendekatan *data augmentation* dapat menjadi tujuan kita, hal ini menyebabkan $L(\beta,\lambda|m)=L(\beta|m)\times L(\lambda|m)$ menafsirkan dengan baik *the induced prior* pada β. Kenyataannya, distribusi posterior dari *augmented data* adalah

$$\pi(\beta,\lambda|\mathbf{m}) \propto \pi(\beta|\mathbf{m})\pi(\lambda)\prod_{k,i,j}\lambda_{kij}^{m_{kij}}, \tag{1.3}$$

dimana $\pi(\lambda)$ adalah distribusi prior bagi λ.

Untuk tahap mendapatkan parameter λ, diasumsikan saling bebas antar $\{\lambda_{kij}, j=0,1\}, \forall k,i.$ Dalam kasus ini, distribusi posterior dengan m tertentu adalah perkalian antara distribusi bagi beta dengan parametemya yang diperbarui oleh **m**.

Bersyarat pada data yang teramati, *the augmented data* memiliki distribusi menurut distribusi binomial yang saling bebas untuk setiap *k*, yaitu

$$\begin{aligned} m_{k01}|\beta,\lambda,n &: \text{Binomial}\left\{n_k, \frac{\lambda_{k01}\theta_0(\vec{x}_i'\beta)}{\sum_i \lambda_{ki1}\theta_i(\vec{x}_i'\beta)}\right\} \\ m_{k10}|\beta,\lambda,n &: \text{Binomial}\left\{N_k-n_k, \frac{\lambda_{k10}\theta_1(\vec{x}_i'\beta)}{\sum_i \lambda_{ki0}\theta_i(\vec{x}_i'\beta)}\right\} \end{aligned} \tag{1.4}$$

Dari konfigurasi ini, sekarang memungkinkan untuk menyimpulkan berdasarkan algoritma *data augmentation* (Tanner, 1996), disebut *the chained data augmentation algorithm* (CDA) yang tahapannya sebagai berikut :

1) Pilih nilai awal β^0 dan λ^0 ;
2) Untuk $i=1,\ldots,t$:
   a) Imputation step
      i) Sampel $\mathbf{m}^i$ dari distribusi binomial yang saling bebas pada (1.6) bersyarat $\beta^{i-1},\lambda^{i-1}$, dan n,
   b) Posterior step
      i) Sampel λ^i dari distribusi beta yang saling bebas bersyarat $\mathbf{m}^i$;
      ii) Sampel β^i dari $\pi(\beta|\mathbf{m})$ pada (1.5) bersyarat $\mathbf{m}^i$.

Kemudian, $\pi(\beta,\lambda|\mathbf{m}^i)$ akan konvergen ke $\pi(\beta,\lambda|\mathbf{n})$ untuk *i* menuju tak hingga (Tunner dan Wong, 1987).

2.3. BAYESIAN MODEL

Dengan memanfaatkan skema data *augmentation* dan menurut Holmes (2003), misalkan *the Bayesian logistic model* sebagai berikut

$$\begin{aligned} y_i &: \text{Bernoulli}\left(g^{-1}(\eta_i)\right) \\ \eta_i &= \mathbf{x}_i\beta \\ \beta &: \pi(\beta) \end{aligned} \tag{1.5}$$

dengan $y_i \in \{0,1\}. k=1,\ldots,N$ adalah variabel respon biner yang dikumpulkan dari *N* objek dengan *p* peubah penjelas $\mathbf{x}_k=(x_{k1},\ldots,x_{kp}), g(u)=\log(u/(1-u))$ adalah fungsi hubung logistik, η_k adalah

penduga linear dan β merupakan vektor koefisien regresi berdimensi $(p \times 1)$ yang prior nya adalah dari distribusi $\pi(g)$.

Model logistik pada (1.5) memiliki kesamaan representasi dengan menggunakan *auxiliary variable*, yaitu

$$\begin{aligned} y_k &= \begin{cases} 1 & ; \ z_k > 0 \\ 0 & ; \ z_k \le 0 \end{cases} \\ z_k &= \mathbf{x}_k\beta + \varepsilon_k \\ \varepsilon_k &: \ \pi(\varepsilon_k) \\ \beta &: \ \pi(\beta) \end{aligned} \tag{1.6}$$

Dalam kasus ini, y_i ditentukan bersyarat pada tanda *auxiliary variable* z_k dan $\pi(\varepsilon_k)$ adalah distribusi logistik baku, dan di bawah kondisi $\varepsilon_k, k = 1,\ldots,N$ saling bebas maka distribusi marginal dari y pada persamaan (1.6) sama dengan model pada persamaan (1.5). Selanjutnya diperkenalkan variabel $\gamma_k, k = 1,\ldots,N$ dan representasi tambahan sebagai berikut:

$$\begin{aligned} y_k &= \begin{cases} 1 & ; \ z_k > 0 \\ 0 & ; \ z_k \le 0 \end{cases} \\ z_k &= \mathbf{x}_k\beta + \varepsilon_k \\ \varepsilon_k &: \ N(0,\gamma_k) \\ \gamma_k &= (2\psi_k)^2 \\ \psi_k &: \ KS \\ \beta &: \ \pi(\beta) \end{aligned} \tag{1.7}$$

dengan $\psi_k, k = 1,\ldots,N$ adalah peubah acak saling bebas yang mengikuti distribusi Kolmogov-Smirnov (Devroye, 1986). Menurut Andrews dan Mallows (1974) dalam kasus ε_k mempunyai skala campuran dari bentuk normal dengan distribusi logistik marginal maka distribusi marginal $\pi(\beta|y)$ untuk model persamaan (1.5), (1.6), dan (1.7) adalah ekivalen.

Keuntungan bekerja dengan persamaan (1.7) adalah bahwa untuk pemilihan $\pi(\beta)$ yang bijaksana, hal itu cocok untuk simulasi yang efisien. Khususnya dalam kasus prior normal pada β, yaitu $\pi(\beta) = N(u,v)$, distribusi bersyarat β adalah tetap normal, yaitu:

$$\begin{aligned} \beta|z,\lambda,y &: \ N(\hat{\beta},\mathbf{V}) \\ \hat{\beta} &= \mathbf{V}(\mathbf{v}^{-1}\mathbf{u} + \mathbf{x}'\mathbf{W}\mathbf{z}) \\ \mathbf{V} &= (\mathbf{v}^{-1} + \mathbf{x}'\mathbf{W}\mathbf{x}). \\ \mathbf{W} &= diag(\gamma_1^{-1},\ldots,\gamma_n^{-1}), \end{aligned} \tag{1.8}$$

dengan $\mathbf{x} = (x'_1,\ldots,x'_N)'$, sedangkan distribusi bersyarat z_k adalah *truncated normal* yang bentuknya sederhana (Robert, 1995), yaitu:

$$z_k \mid \beta, x_k, y_k, \lambda_k \propto \begin{cases} N(x_k\beta,\gamma_k)I(z_k > 0) & ; \ y_k = 1 \\ N(x_k\beta,\gamma_k)I(z_k \le 0) & ; \ \text{lainnya} \end{cases} \tag{1.9}$$

Hal yang sama juga berlaku untuk *the Bayesian probit model*, dalam konstruksi/skema data *augmentation*.

2.4. Simulasi data menggunakan model Bayes

Misalkan $Y_1,\ldots,Y_N$ adalah peubah acak biner yang saling bebas, Y_k : *Bernoulli*$(p(Y_k=1)=\phi_k)$ dan ϕ_k terkait dengan peubah penjelas $\mathbf{x}_k=(x_{k1},\ldots,x_{kp})'$ melalui regresi binomial dengan kesalahan klasifikasi. Model respon biner $p_k=\Psi(x'_k\beta)$ dan g(g)=Ψ^{-1} adalah fungsi hubung. Model kesalahan klasifikasi ditulis dalam bentuk:

$$\phi_k = p_k(1-\lambda_{10})+(1-p_k)\lambda_{01} \tag{1.10}$$

Dengan p_k adalah peluang benar positif untuk pengamatan ke-*k*, λ_{10} adalah peluang salah negatif, dan λ_{01} adalah peluang salah positif.

Berikut dikenalkan peubah laten pada skema data *augmentation* yaitu $c_{ij}^k, i,j=0,1$, dimana $c_{11}^k=1$ jika *k* adalah benar positif, dimana $c_{10}^k=1$ jika *k* adalah salah negatif, dimana $c_{01}^k=1$ jika *k* adalah salah positif, dan dimana $c_{00}^k=1$ jika *k* adalah benar negatif. Sehingga setiap unit pengamatan mempunyai vektor laten $\mathbf{c}^k=(c_{11}^k,c_{10}^k,c_{01}^k,c_{00}^k)$. Fungsi likelihood dapat ditulis:

$$L(\beta,\lambda|D)\propto\prod_{k=1}^{N}\left[\{p_k(1-\lambda_{10})+(1-p_k)\lambda_{01}\}^{y_k}\{p_k\lambda_{10}+(1-p_k)(1-\lambda_{01})\}^{1-y_k}\right] \tag{1.11}$$

Misalkan distribusi prior dari β adalah $N(\mathbf{b}_0,\mathbf{B}_0)$, dengan diberikan data D, distribusi posterior bersama bagi *unobservables* c,β, dan λ adalah

$$\begin{aligned}\pi(c,\beta,\lambda|D)\propto{}&\pi(\beta)\pi(\lambda)\\ &\times\prod_{k=1}^{N}\left[\{p_k(1-\lambda_{10})\}^{c_{11}^k}\{p_k\lambda_{10}\}^{c_{10}^k}\{(1-p_k)\lambda_{01}\}^{c_{01}^k}\{(1-p_k)(1-\lambda_{01})\}^{c_{00}^k}\right]\\ &\times\left(I[y_k=1]I[c_{11}^k+c_{01}^k=1]+I[y_k=0]I[c_{10}^k+c_{00}^k=1]\right)\end{aligned} \tag{1.12}$$

Untuk dapat menjawab tujuan pada tulisan ini, maka dirancang skema simulasi sebagai berikut. Simulasi untuk membangkitkan data didasarkan model di atas, dengan langkah-langkah sebagai berikut :

i. Ditentukan dua peubah penjelas x_{k1} dan x_{k2} dibangkitkan dengan x_{k1} : $N(2,0.09)$ dan x_{k2} : $N(3,0.09)$, untuk dua kasus yaitu kasus ukuran sampel besar: $k=1,\ldots,100$ dan kasus ukuran sampel kecil : $k=1,\ldots,20$.

ii. Nilai peluang didapatkan dengan model

iii. $\eta_k=\Psi^{-1}(p_k)=\beta_0+\beta_1x_{k1}+\beta_2x_{k2}$

dengan menggunakan *link function* logit dan probit, maka didapat nilai peluang p_k, dan ditentukan spesifikasi prior untuk parameter regresi yaitu $\beta=(\beta_0,\beta_1,\beta_2)=(2,-4,2)$.

Untuk setiap model, *the true binary dependent variable* $y^{true}=y^T$ didapat dari

$$y^T=\begin{cases}1 & : \ p_k>0.5\\ 0 & : \ p_k\le 0.5\end{cases}$$

iv. Beberapa hasil pada langkah (iii), ditetapkan sebagai salah klasifikasi yaitu

Untuk kasus ukuran sampel besar

7 nilai dengan $y^T=0$ menjadi $y^T=1$,

5 nilai dengan $y^T = 1$ menjadi $y^T = 0$.

Untuk kasus ukuran sampel kecil

2 nilai dengan $y^T = 0$ menjadi $y^T = 1$,

3 nilai dengan $y^T = 1$ menjadi $y^T = 0$.

v. Selanjutnya didapat variabel baru y yang mengandung kesalahan klasifikasi yaitu y^* menggantikan y^T dan dapat ditentukan variabel laternya.

vi. Kemudian diketahui proporsi kesalahan klasifikasi untuk variabel y^*, diberikan oleh
Untuk kasus (1), misalnya

$$\lambda_{01} = p(\text{salah positif}) = 1 - \text{specificity} = \frac{7}{45}$$

$$\lambda_{10} = p(\text{salah negatif}) = 1 - \text{sensitivity} = \frac{5}{55}$$

Untuk kasus (2), misalnya

$$\lambda_{01} = p(\text{salah positif}) = 1 - \text{specificity} = \frac{2}{8}$$

$$\lambda_{10} = p(\text{salah negatif}) = 1 - \text{sensitivity} = \frac{3}{12}$$

vii. Dua kasus kesalahan klasifikasi dicobakan pada kedua model, yaitu
1) Hasil y^T yang dekat dengan border, $p_k \approx 0.5$, ditetapkan sebagai y salah klasifikasi.
2) Hasil y^T yang jauh dengan border, $p_k \approx 0$ atau $p_k \approx 1$, ditetapkan sebagai y salah klasifikasi.

viii. Analisis data : menduga salah klasifikasi, menduga parameter regresi dan ukuran kesesuaian model. Ulangan dilakukan sebanyak 200 kali untuk kasus ukuran sampel besar, dan ulangan dilakukan sebanyak 50 kali untuk kasus ukuran contoh kecil.

3. HASIL DAN PEMBAHASAN

Setelah data hasil simulasi didapatkan, dilakukan analisis data untuk menduga salah klasifikasi, menduga parameter regresinya, dan ukuran kesesuaian model (menggunakan AIC). Hasil lengkapnya diberikan pada tabel berikut.

Tabel 1. Kesalahan klasifikasi dugaan dan peluang kesalahan klasifikasi dugaan

Fungsi hubung (*link function*)	Ukuran sampel	Kesalahan klasifikasi		Peluang kesalahan klasifikasi		AIC
		y^{01}	y^{10}	λ_{01}	λ_{10}	
Kasus 1 : $p \approx 0.5$						
Logit	N = 100	8 (7)	7 (5)	8/50 (7/53)	7/50 (5/47)	38.6206
Probit		3 (7)	8 (5)	3/42 (7/55)	8/58 (5/45)	69.2256
Logit	N = 20	0 (2)	6 (3)	0 (2/15)	6/20 (3/5)	15.7033
Probit		2 (2)	3 (3)	2/9 (2/11)	3/11 (3/9)	18.2145
Kasus 2 : $p \approx 0$ atau $p \approx 1$						
Logit	N = 100	15 (7)	7 (5)	15/63 (7/47)	7/37 (5/53)	58.4184
Probit		19 (7)	7 (5)	19/31 (7/45)	7/69 (5/55)	92.5873
Logit	N = 20	6 (2)	0 (3)	6/20 (2/7)	0 (3/13)	17.2404
Probit		5 (2)	5 (3)	5/10 (2/11)	5/10 (3/9)	27.2149

Keterangan : Angka pada kurung (.) menunjukkan angka salah klasifikasi yang ditetapkan dalam simulasi

Tabel 1 menunjukkan bahwa model yang dapat menduga kesalahan klasifikasi dengan lebih baik (lebih mendekati kesalahan klasifikasi yang ditetapkan dalam simulasi) adalah model logistik, untuk kasus (1) yaitu kasus dimana individu memiliki $p \approx 0.5$. Tetapi jika individu yang menjadi salah klasifikasi adalah individu dengan $p \approx 0$ atau $p \approx 1$, maka baik model probit maupun logistik tidak dapat menduga peluang salah klasifikasi dengan baik (dugaan jauh dari yang ditetapkan dalam simulasi). Hal ini disebabkan oleh individu pencilan (nilai $p \approx 0$ atau $p \approx 1$) menjadi individu tidak pencilan setelah dibuat salah klasifikasi, akibatnya akan sangat mempengaruhi hasil dugaan.

Dari Tabel 1 juga dapat disimpulkan bahwa model logistik merupakan model lebih baik dibandingkan dengan model probit, karena secara umum nilai AIC model logistik lebih kecil daripada nilai AIC model probit.

Tabel 2. Rata-rata dan deviasi standar dugaan parameter

Fungsi hubung (*link function*)	Ukuran sampel	Rata-rata			Deviasi standar		
		$\hat{\beta}_0$ (2)	$\hat{\beta}_1$ (-4)	$\hat{\beta}_2$ (2)	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
Kasus 1 : $p \approx 0.5$							
Logit	N=100	2.4841	-4.4988	2.2107	3.2344	0.6846	0.8548
Probit		1.4014	-2.3513	1.1391	1.6112	0.3891	0.4385
Logit	N=20	2.0099	-3.4463	1.6751	1.8647	0.4854	0.5663
Probit		4.1117	-7.9719	3.9753	2.7525	0.8465	0.8922
Kasus 2 : $p \approx 0$ atau $p \approx 1$							
Logit	N=100	1.2481	-2.2683	1.1214	2.4904	0.4579	0.6303
Probit		1.1846	-2.2195	1.1157	2.7884	0.4942	0.6944
Logit	N=20	1.7364	-5.2727	2.9528	2.3753	0.5105	0.7259
Probit		1.3887	-2.4779	1.1959	1.9041	0.3757	0.5133

Keterangan : Angka pada kurung (.) menunjukkan nilai parameter yang ditetapkan dalam simulasi

Tabel 2 mendukung rangkuman hasil analisis pada Tabel 1 yaitu bahwa, model logistik dapat menduga parameter regresi $\left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\right)$ dengan lebih baik (nilai dugaan mendekati nilai parameter yang ditetapkan dalam simulasi) dibandingkan dengan model probit baik untuk ukuran sampel besar maupun untuk ukuran sampel kecil, untuk kasus (1). Tetapi pada kasus (2), tidak satupun model probit maupun logistik yang dugaan parameternya mendekati nilai sebenarnya (ditetapkan dalam simulasi). Berdasarkan Tabel 1 dan Tabel 2, secara umum kesalahan klasifikasi menyebabkan pendugaan parameter menjadi tidak valid. Terutama, nilai dugaan parameter menjadi jauh dari nilai yang sebenarnya jika kesalahan klasifikasi terjadi pada peubah respon yang jauh dari nilai border (kasus 2), artinya jika yang menjadi salah klasifikasi adalah data pencilan, maka efeknya ke pendugaan lebih signifikan.

4. KESIMPULAN

Kesalahan klasifikasi menyebabkan pendugaan parameter menjadi tidak valid karena kesalahan klasifikasi membuat peluang terjadi sukses menjadi lebih besar dari yang sebenarnya, terutama jika kesalahan klasifikasi terjadi pada peubah respon dari individu yang jauh dari border, yaitu individu yang menjadi pencilan.

Ukuran sampel besar maupun kecil memiliki pengaruh yang sama bagi hasil dugaan parameter, artinya jika ukuran sampel besar menyebabkan dugaan parameter tidak valid, maka ukuran sampel kecil memberikan kesimpulan yang sama, begitupun sebaliknya.

DAFTAR PUSTAKA

Bedrick EJ, Christensen R, dan Johnson W. (1996). A new perspective on priors for generalized linear models. *Journal of the American Statistical Association* 91, 1450-1460.

Chen Z, Yi GY, and Wu C. (2011). Marginal methods for correlated binary data with misclassified responses. *Biometrika*, 98,3, pp. 647-662.

Paulino CD, Silva G, Achcar JA. (2005). Bayesian analysis of correlated misclassified binary data. *Computational Statistics and Data Analysis* **49**, 1120-1131.

Paulino CD, Soares P, Neuhaus J. (2003). Binomial Regression with Misclassification. *Biometrics*, vol 59, pp. 670-675.

McCullagh P, and Nelder JA. (1989). *Generalized Linear Models*. 2nd *Edition*, Chapman and Hall, New York.

Holmes CC, dan Knorr-Held L. (2003). Efficient simulation of Bayesian logistic regression models. http://epub.ub.uni-muenchen.de/

ISBN : 978 – 979 – 16353 – 9 – 4

PROSIDING

SEMINAR NASIONAL
MATEMATIKA DAN PENDIDIKAN MATEMATIKA

"Penguatan Peran Matematika dan Pendidikan Matematika
Untuk Indonesia yang Lebih Baik"

Penyelenggara :
Jurusan Pendidikan Matematika
FMIPA UNY

Yogyakarta, 9 November 2013

PROSIDING

Prosiding dapat diakses:
http://eprints.uny.ac.id/view/subjects/snmpm2013.html