Estimasi Parameter Model Logit pada Respons Biner Multivariat Menggunakan Metode Mle dan Gee

Estimating Parameters of Logit Model on Multivariate Binary Response Using Mle and Gee

Jaka Nugraha[1)], Suryo Guritno[2)], Sri Haryatmi[2)]
[1]*Jurusan Statistika UII,* [2]*Jurusan Matematika UGM*

ABSTRACT

In this paper, we discuss binary multivariate response modeling based on extreme value distribution. Independent variables used in these models are some attributes of the alternative (labeled Z_{ijt}) and some attributes of the decision maker (labeled X_i). We assumed that n the decision maker observed with T response. Y_{it} is t^{nd} response variables from decision maker i and value Y_{it} is binary. Response of decision maker i can be expressed as $Y_i = (Y_{i1},...,Y_{iT})$. In each of the decision maker, we have data (Y_i, X_i, Z_i). Models are derived by the assumption that maximum random utility which the decision maker i choose one of the alternatives having greatest utility. Methods of parameter estimation are Maximum Likelihood Estimator (MLE) method and Generalized Estimating Equation (GEE). First discussion in this study is the estimation by MLE with independent assumption among response and then the MLE estimation using joint distribution by Bahadur's representation. By MLE and GEE, estimating equations are obtained and solved by numerical (like's Newthon-Rahpson method) in the condition that not all of the parameters of individual attributes can be estimated (identified). Based on testing simulation data with R.2.5.0, we recommend (a) in low correlation, GEE is better than MLE (b) in moderate correlation, MLE is most efficient but not stable (c) in high or moderate correlation, MLE and GEE should be used (d) correlation estimators cannot explain the real correlation because of its bias.

Keywords : Random utility models, logit models, discrete choice models

PENDAHULUAN

Model pemilihan diskrit menggambarkan pembuat keputusan memilih diantara alternatif yang tersedia. Model pemilihan diskrit diturunkan dengan asumsi bahwa pembuat keputusan memilih alternatif yang mempunyai utiliti terbesar (Train 2003). Model regresi yang diturunkan dari distribusi nilai ektreme merupakan bagian dari pemodelan Pilihan Diskret (DCM, *Discrete Choice Model*). Model rccccegresi pada respon nominal yang diperoleh dcengan menggunakan distribusi nilai ektreme cadalah sama dengan model logistik (Nugraha *et al*. 2006).

Liang & Zeger (1986) menyampaikan bahwa analisis logistik maupun probit pada data panel dengan menggunakan pendekatan univariat, yakni mengabaikan adanya korelasi akan menghasilkan estimator parameter yang masih konsinten tetapi jika terdapat korelasi yang besar maka penaksir tersebut menjadi tidak efisien. Prentice (1988) menyampaikan strategi pemodelan menggunakan pendekatan GEE untuk mendapatkan estimator koefisien regresi yang konsisten dan asimtotis normal. Pendekatan GEE tidak menggunakan perhitungan integral rangkap.

Contoyanis *et al.* (2002) menyatakan bahwa pada model efccek tetap jika terdapat korelasi antara efcccek individu terhadap variabel independen maka estimator yang diperoleh menjadi tidak efisien. Dengan mengasumsikan model efek random, estimator yang didapatkan menjadi lebih efisien. Harris *et al.* (2000) telah melakukan pengujian sifat-sifat estimator model probit pada data panel. Disimpulkan bahwa meskipun jumlah sampel terbatas, estimasi menggunakan MLE masih menghasilkan estimator yang baik. MLE mempunyai sifat-sifat yang baik untuk sampel besar, khususnya *asymptotically efficient* (Horrowitz *et al.* 2001).

Dalam makalah ini akan dibahas estimasi parameter model logit pada respon biner multivariat. Variabel independen yang

dipertimbangkan dalam model adalah variabel karakteristik individu dan variabel karakteristik pilihan. Metode estimasi parameter menggunakan MLE dan metode GEE. Pembahasan estimasi MLE diawali dengan asumsi independen antar respon dan kemudian dilanjutkan pembahasan estimasi MLE menggunakan distribusi gabungan dari *Bahadur's representation*.

Beberapa metode estimasi tersebut diimplentasikan pada data simulasi. Data dibangkitkan pada nilai parameter yang sudah ditentukan. Hasil estimasi dibandingkan dengan nilai parameter sesungguhnya.

Model utilitas

Diasumsikan bahwa n individu masing-masing diobservasi sebanyak T respon. Y_{it} adalah respon ke-t pada individu/subjek ke-i dan setiap responnya adalah biner. Sehingga respon pada individu ke-i, dapat disajikan dalam bentuk $\mathbf{Y}_i = (Y_{i1},....,Y_{iT})$ sebagai vektor 1xT, Setiap individu mempunyai kovariate X_i sebagai karakteristik individu i dan kovariate Z_{ijt} sebagai karakteristik alternatif/pilihan j pada individu i. Untuk menyederhanakan penulisan, diambil satu variabel karakteristik individu dan satu variabel karakteristik pilihan. Utilitas subjek i memilih alternatif j pada periode t adalah

$$U_{ijt} = V_{ijt} + \varepsilon_{ijt}$$
untuk t=1,2,...,T ; i=1,2,...,n ; j=0,1 (1)
$$V_{ijt} = \alpha_{jt} + \beta_{jt}X_i + \gamma_t Z_{ijt}$$

U_{ijt} adalah utilitas yang merupakan variabel laten dan V_{ijt} dinamakan representatif utilitas. Dengan mengasumsikan bahwa pembuat keputusan (subjek) menentukan pilihan berdasarkan nilai utilitas yang maksimum, maka model dapat disajikan dalam bentuk selisih utilitas,

$$U_{it} = U_{i1t} - U_{i0t} = (V_{i1t} - V_{i0t}) + (\varepsilon_{i1t} - \varepsilon_{i0t}) = V_{it} + \varepsilon_{it} \quad (2)$$

Selanjutnya dapat disusun hubungan antara Y_{it} dan variabel laten U_{it}, yaitu

$$y_{it} = 1 <=> U_{i1t} > U_{i0t} <=> U_{it} > 0 <=> -V_{it} < \varepsilon_{it}$$

Probabilitas subjek i memilih ($y_{i1} = 1,...., y_{iT} = 1$) adalah

$P(y_{i1} = 1,...., y_{iT} = 1) = P(0 < U_{i1},..., 0 < U_{iT}) = P(-V_{i1} < \varepsilon_{i1},..., -V_{iT} < \varepsilon_{iT}) =$

$$\int_{\varepsilon i} I(-V_{it} < \varepsilon_{it}).f(\varepsilon_i^{'})d\varepsilon_i^{'} \qquad \forall t \qquad (3)$$

dengan $\varepsilon_i = (\varepsilon_{i1},..., \varepsilon_{iT})$'. Nilai probabilitas ini merupakan hitungan integral rangkap T dan tergantung pada parameter α, β dan γ maupun distribusi ε.

MLE pada model logit

Model Logit dapat diturunkan dari asumsi bahwa ε_{ijt} berdistribusi nilai ekstrim (*extreme value*) yang saling independen untuk semua i,j dan t. Fungsi densitas nilai ekstrem (Gumbel) adalah

$$f(\varepsilon_{ijt}) = \varepsilon^{-\varepsilon_{ijt}}\varepsilon^{-\varepsilon^{-\varepsilon_{ijt}}} \qquad (4)$$

Teorema 1

Jika ε_{ijt} adalah variabel random yang mempunyai densitas *extreme value* maka berdasarkan asumsi bahwa pembuat keputusan (subjek) menentukan pilihan berdasarkan nilai utilitas yang maksimum, probabilitas subjek i memilih j=1 untuk respon ke-t (probabilitas marginal) adalah

$$P(y_{it} = 1) = \pi_{it} = \frac{\exp(V_{i1t})}{[\exp(V_{i0t}) + \exp(V_{i1t})]}$$

dengan $V_{ijt} = \alpha_{jt} + \beta_{jt}X_i + \gamma_t Z_{ijt}$ untuk t=1,2,...,T ; i=1,2,...,n ; j=0,1.

Jika variabel random ε_{ijt} saling independen untuk semua i,j dan t maka fungsi log-likelihoonya adalah

$$LL = \sum_{i=1}^{n}\sum_{t}^{T}\{y_{it}(V_{i1t} - V_{i0t}) + V_{i0t} - \ln([\exp(V_{i0t}) + \exp(V_{i1t})])\}$$

(5)

Proposisi 1.

Derivarif pertama fungsi loglikelihood (5) adalah

$$\frac{\partial LL_t}{\partial \alpha_{0t}} = \sum_{i=1}^{n}(\pi_{it} - y_{it});$$

$$\frac{\partial LL_t}{\partial \alpha_{1t}} = \sum_{i=1}^{n}(y_{it} - \pi_{it});$$

$$\frac{\partial LL_t}{\partial \beta_{1t}} = \sum_{i=1}^{n} X_i(\pi_{it} - y_{it})$$

$$\frac{\partial LL_t}{\partial \beta_{1t}} = \sum_{i=1}^{n} X_i \left(y_{it} - \pi_{it}\right) \quad \text{dan}$$

$$\frac{\partial LL_t}{\partial \gamma_t} = \sum_{i=1}^{n} (Z_{i1t} - Z_{i0t})\left(y_{it} - \pi_{it}\right)$$

untuk t=1,2,..,T.

Pada kondisi derivatif pertama sama dengan nul maka agar MLE untuk parameter α dan β teridentifikasi/terestimasi maka perlu disyaratkan α_{0t} =1 dan β_{0t}=1. Sehingga parameter yang harus diestimasi adalah θ_t = (α_t, β_t, γ_t)' dengan α_t=α_{1t} dan β_t=β_{1t}. Fungsi log-likehood (5) menjadi

$$\frac{\partial LL}{\partial \theta_t} = \sum_{i=1}^{n} \begin{pmatrix} 1 \\ X_i \\ (Z_{i1t} - Z_{i0t}) \end{pmatrix} \left(y_{it} - \pi_{it}\right) \qquad (6)$$

Proposisi 2.
Derivarif ke-dua fungsi loglikelihood (5) pada α_{0t} =1 dan β_{0t}=1 adalah

$$\frac{\partial LL}{\partial \theta_t' \partial \theta_t} = -\sum_{i=1}^{n} \begin{pmatrix} 1 & X_i & (Z_{i1t} - Z_{i0t}) \\ X_i & X_i^2 & X_i(Z_{i1t} - Z_{i0t}) \\ (Z_{i1t} - Z_{i0t}) & X_i(Z_{i1t} - Z_{i0t}) & (Z_{i1t} - Z_{i0t})^2 \end{pmatrix} \pi_{it}(1-\pi_{it})$$

dengan θ_t = (α_t, β_t , γ_t)'dan merupakan matrik definet negatif.

Teorema 2.
MLE untuk parameter θ_t = (α_t, β_t , γ_t)' dengan asumsi independen antar respon Y_{it} adalah penyelesaian dari persamaan

$$\sum_{i=1}^{n} \begin{pmatrix} 1 \\ X_i \\ (Z_{i1t} - Z_{i0t}) \end{pmatrix} \left(y_{it} - \pi_{it}\right) = 0$$

yang berlaku untuk setiap t=1,..,T dengan α_{0t}=1 dan β_{0t} =1

Lemma 1.
Penyelesaian persamaan penaksir pada Teorema 2 untuk parameter θ_t = (α_t, β_t , γ_t) dan t tertentu dengan menggunakan metode Newton-Raphson pada iterasi ke-(k+1) adalah

$$\theta_t^{(k+1)} = \theta_t^{(k)} - (H_t^{(k)})^{-1} g_t^{(k)}$$

dan penaksir MLE parameter θ mempunyai sifat

$$\hat{\theta}_t \xrightarrow{a} N[\theta_t, \{H_t\}^{-1}]$$

dengan $g_t^{(k)} = \sum_{i=1}^{n} \begin{pmatrix} 1 \\ X_i \\ (Z_{i1t} - Z_{i0t}) \end{pmatrix} \left(y_{it} - \pi_{it}^{(k)}\right)$

dan

$$H_t^{(k)} = -\sum_{i=1}^{n} \begin{pmatrix} 1 & X_i & (Z_{i1t} - Z_{i0t}) \\ X_i & X_i^2 & X_i(Z_{i1t} - Z_{i0t}) \\ (Z_{i1t} - Z_{i0t}) & X_i(Z_{i1t} - Z_{i0t}) & (Z_{i1t} - Z_{i0t})^2 \end{pmatrix} \pi_{it}(1-\pi_{it})$$

MLE menggunakan *Bahadur's Represetation*
Penyajian distribusi bersama untuk T respon biner, $Y_{i1},....,Y_{iT}$ dapat dituliskan sebagai
$P(Y_{i1}=y_{i1} ,...,Y_{iT}= y_{iT})$=

$$\prod_{t=1}^{T} \pi_{it}^{y_{it}} (1-\pi_{it})^{1-y_{it}} \left(1 + \sum_{t<k} \rho_{tk} e_{it} e_{ik} + \sum_{t<k<l} \rho_{tkl} e_{it} e_{ik} e_{il} + ... + \rho_{12..T} e_{i1} e_{i2} .. e_{iT}\right) \qquad (7)$$

dengan $e_{it} = \dfrac{(Y_{it} - \pi_{it})}{[\pi_{it}(1-\pi_{it})]^{1/2}}$ dan ρ_{tk} = $E(e_{it}e_{ik}),...., \rho_{12...T} = E(e_{i1}e_{i2}...e_{iT})$.

Penulisan distribusi pada persamaan (7) disebut *Bahadur's representation* (Fitzmaurice *et al*. 1993) yang memuat parameter korelasi ρ = (ρ_{12}, ρ_{13},..., $\rho_{123...T}$).

Jika diasumsikan bahwa korelasi tiga respon atau lebih bernilai nul, maka persamaan (7) dapat disederhanakan menjadi

$P(Y_{i1}=y_{i1} ,...,Y_{iT}= y_{iT})$=

$$\prod_{t=1}^{T} \pi_{it}^{y_{it}} (1-\pi_{it})^{1-y_{it}} \left(1 + \sum_{t<k} \rho_{tk} \frac{(Y_{it} - \pi_{it})(Y_{ik} - \pi_{ik})}{[\pi_{it}(1-\pi_{it})\pi_{ik}(1-\pi_{ik})]^{1/2}}\right) \qquad (8)$$

dengan parameter korelasi ρ = (ρ_{12}, ρ_{13},..., $\rho_{(T-1)T}$).

Parameter ρ merepresentasikan ukuran assosiasi antar respon. Asosiasi (hubungan) antara Y_{it} dan Y_{ik} dapat berupa ukuran koefisien korelasi, rasio odd (*odds ratio)* ataupun risiko relatif *(relative risk)*. Koefisien korelasi ρ_{itk} adalah

$$\rho_{itk} = \frac{\mathrm{Cov}(Y_{it}, Y_{ik})}{\sqrt{\mathrm{var}(Y_{it})}\sqrt{\mathrm{var}(Y_{ik})}}$$

$$= \frac{\pi_{itk} - \pi_{it}\pi_{ik}}{[\pi_{it}(1-\pi_{it})\pi_{ik}(1-\pi_{ik})]^{1/2}} \quad (8)$$

dengan $\pi_{itk} = P(Y_{it}=1, Y_{ik}=1)$.

Selanjutnya parameter ρ_{itk} akan diestimasi dari distribusi bersama $(Y_{it}=y_{it}; Y_{ik}=y_{ik})$. Dari persamaan (9) diperoleh

$P(Y_{it}=1, Y_{ik}=1)=$

$$\pi_{itk} = \pi_{it}\pi_{ik} + \rho_{tk}[\pi_{it}(1-\pi_{it})\pi_{ik}(1-\pi_{ik})]^{1/2}$$

$P(Y_{it}=1, Y_{ik}=0)$

$$= \pi_{it}(1-\pi_{ik}) - \rho_{tk}[\pi_{it}(1-\pi_{it})\pi_{ik}(1-\pi_{ik})]^{1/2}$$

$P(Y_{it}=0, Y_{ik}=1)$

$$= \pi_{ik}(1-\pi_{it}) - \rho_{tk}[\pi_{it}(1-\pi_{it})\pi_{ik}(1-\pi_{ik})]^{1/2}$$

$P(Y_{it}=0, Y_{ik}=0)$

$$= (1-\pi_{it})(1-\pi_{ik}) - \rho_{tk}[\pi_{it}(1-\pi_{it})\pi_{ik}(1-\pi_{ik})]^{1/2}$$

Dari (a), (b), (c) dan (d), fungsi probabilitas bivariat $P(Y_{it}, Y_{ik})$ dapat dituliskan sebagai

$$P(Y_{it}=y_{it}, Y_{ik}=y_{ik}) =$$
$$\pi_{it}^{y_{it}}(1-\pi_{it})^{1-y_{it}}\pi_{ik}^{y_{ik}}(1-\pi_{ik})^{1-y_{ik}} + (-1)^{y_{it}+y_{ik}}\rho_{tk}[\pi_{it}(1-\pi_{it})\pi_{ik}(1-\pi_{ik})]^{1/2} \quad (9)$$

Estimator tak bias dari ρ_{tk} adalah

$$\hat{\rho}_{tk} = \frac{(Y_{it}-\hat{\pi}_{it})(Y_{ik}-\hat{\pi}_{ik})}{[\hat{\pi}_{it}(1-\hat{\pi}_{it})\hat{\pi}_{ik}(1-\hat{\pi}_{ik})]^{1/2}} \quad (10)$$

Proposisi 3.
Parameter ρ_{tk} yang merupakan korelasi antara Y_t dan Y_k mempunyai nilai

$$-\min\left\{\left(\frac{\pi_{it}\pi_{ik}}{(1-\pi_{it})(1-\pi_{ik})}\right)^{1/2}; \left(\frac{(1-\pi_{it})(1-\pi_{ik})}{\pi_{it}\pi_{ik}}\right)^{1/2}\right\} \le \rho_{tk} \le$$

$$\min\left\{\left(\frac{(1-\pi_{it})\pi_{ik}}{\pi_{it}(1-\pi_{ik})}\right)^{1/2}, \left(\frac{\pi_{it}(1-\pi_{ik})}{(1-\pi_{it})\pi_{ik}}\right)^{1/2}\right\}$$

Proposisi 4.
Berdasarkan *Bahadur's representation* dan mengabaikan korelasi tiga respon atau lebih, maka fungsi log-likelihood untuk parameter $\theta_t = (\beta_t;\gamma_t)'$ adalah

$$LL(\theta) = \sum_{i=1}^{n}\sum_{t=1}^{T} \ln\left(\pi_{it}^{y_{it}}(1-\pi_{it})^{1-y_{it}}\right) + \sum_{i=1}^{n}\sum_{t=1}^{T} \ln\left(1 + \sum_{t<k}\rho_{tk}\frac{(Y_{it}-\pi_{it})(Y_{ik}-\pi_{ik})}{[\pi_{it}(1-\pi_{it})\pi_{ik}(1-\pi_{ik})]^{1/2}}\right)$$

$$\frac{\partial^2 LL_t}{\partial\theta_t'\partial\theta_t} = -\sum_{i=1}^{n}\begin{pmatrix} 1 & X_i & (Z_{i1t}-Z_{i0t}) \\ X_i & X_i^2 & X_i(Z_{i1t}-Z_{i0t}) \\ (Z_{i1t}-Z_{i0t}) & X_i(Z_{i1t}-Z_{i0t}) & (Z_{i1t}-Z_{i0t})^2 \end{pmatrix} \left(\pi_{it}(1-\pi_{it}) + (1-C)C\right)$$

$$B_{itk} = \frac{(Y_{it}-\pi_{it})(Y_{ik}-\pi_{ik})}{[\pi_{it}(1-\pi_{it})\pi_{ik}(1-\pi_{ik})]^{1/2}}$$

$$\frac{\partial LL_t}{\partial\theta_t} = \sum_{i=1}^{n}\begin{pmatrix} X_i \\ (Z_{i1t}-Z_{i0t}) \end{pmatrix}\left(y_{it}-\pi_{it} + \left(1+\sum_{t<k}\rho_{tk}B_{itk}\right)^{-1} \sum_{t<k}\rho_{tk}\frac{\left(2\pi_{it}(1-\pi_{it})-\frac{Y_{ik}}{2}\right)(Y_{ik}-\pi_{ik})}{[\pi_{it}(1-\pi_{it})]^{1/2}[\pi_{ik}(1-\pi_{ik})]^{1/2}}\right)$$

dengan

$$C = \left(1+\sum_{t<k}\rho_{tk}B_{itk}\right)^{-1}\sum_{t<k}\rho_{tk}(1-2\pi_{it})(Y_{ik}-\pi_{ik})^2 \left(1+\frac{Y_{ik}}{4}[\pi_{it}(1-\pi_{it})]^{-1}\right)$$

Proposisi 5.
Persamaan penaksir untuk parameter θ_t adalah

$$\sum_{i=1}^{n}\begin{pmatrix} 1 \\ X_i \\ (Z_{i1t}-Z_{i0t}) \end{pmatrix}\left(y_{it}-\pi_{it}+\Psi_{it}\right) = 0 \text{ untuk}$$

t=1,...,T

GEE untuk model logit
Liang & Zeger (1986) dan Prentice (1988) telah mengembangkan GEE. Pendekatan GEE menghasilkan estimator konsisten untuk parameter regresi, di bawah spesifikasi yang benar untuk fungsi mean, π_i yang merupakan vektor respons untuk masing-masing individu. GEE menggunakan pendekatan distribusi marginal. GEE untuk $\boldsymbol{\theta}$ dapat dituliskan dalam bentuk

$$G(\boldsymbol{\theta}) = \sum_{i=1}^{n} W_i \Delta_i S_i^{-1} (Y'_i - \pi'_i) = 0 \qquad (11)$$

$$\Delta_i = diag\left(\pi_{i1}(1-\pi_{i1}) \quad \ldots \quad \pi_{iT}(1-\pi_{iT})\right)$$

$$W_i = diag\left(\begin{pmatrix} 1 \\ X_i \\ (Z_{i11} - Z_{i01}) \end{pmatrix}, \ldots, \begin{pmatrix} 1 \\ X_i \\ (Z_{i1T} - Z_{i0T}) \end{pmatrix}\right)$$

dan

$\mathbf{Y}_i = (Y_{i1},\ldots,Y_{iT})$; $\boldsymbol{\pi}_i = (\pi_{i1},\ldots,\pi_{iT})$; $\mathbf{S}_i = \mathbf{A}_i^{1/2}\mathbf{R}_i\mathbf{A}_i^{1/2}$;

$$\mathbf{A}_i^{1/2} = diag\left(\sqrt{Var(Y_{i1})} \quad \ldots \quad \sqrt{Var(Y_{iT})}\right)$$

$\mathbf{R}_i$ adalah matrik *"working"* korelasi $\mathbf{Y}_i$ dan $\mathbf{W}_i$ disebut matrik observasi.

Untuk mengestimasi R_i, didefinisikan vektor korelasi empirik r_i yang berukuran T(T-1)/2 dengan elemen-elemen

$$r_{ist} = \frac{(Y_{is} - \pi_{is})(Y_{it} - \pi_{it})}{[\pi_{is}(1-\pi_{is})\pi_{it}(1-\pi_{it})]^{1/2}} \qquad (12)$$

Korelasi empirik r_{ist} merupakan estimator tak bias parameter ρ_{ist} untuk i = 1,2,..n dan s,r=1,2,…T. Jika diasumsikan $\rho_{ist} = \rho_{st}$ untuk semua i maka penaksirnya adalah

$$\hat{\rho}_{st} = \frac{1}{n}\sum_{i=1}^{n} r_{ist} \qquad (13)$$

Persamaan (12) dan (14) diselesaikan secara bersamaan untuk mendapatkan penaksir parameter θ dan ρ. Prosedur mendapatkan penaksir θ dan ρ adalah memberikan nilai penduga awal untuk R_i untuk menghitung parameter θ secara iterasi Newton-Raphson. Nilai θ yang diperoleh pada langkah (1) digunakan untuk menghitung ρ (rumus pada persamaan (13) dan (14)), nilai ρ dari langkah (2) untuk menghitung paramater θ menggunakan iterasi Newton-Raphson. Langkah (2) dan (3) diulang sampai mencapai titik konvergen.

Persamaan iterasi Newton-Raphson ke (k+1) untuk parameter θ adalah $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} -$

$$\left(\sum_{i=1}^{n} W_i \Delta_i S_i^{-1} \Delta_i W_i^{'}\right)^{-1} \left(\sum_{i=1}^{n} W_i \Delta_i S_i^{-1} (Y'_i - \pi'^{(k)}_i)\right) \qquad (14)$$

Liang & Zeger (1986) merekomendasikan estimator yang konsisten untuk Var($\boldsymbol{\theta}$),

$$\hat{Var}(\hat{\theta}) = M_0^{-1} M_1 M_0^{-1} \qquad (15)$$

$$M_0 = \sum_{i=1}^{n}\left(\frac{\partial \hat{\pi}_i}{\partial \theta}\right) S_i^{-1} \left(\frac{\partial \hat{\pi}_i}{\partial \theta}\right)^1 \quad \text{dan}$$

$$M_1 = \sum_{i=1}^{n}\left(\frac{\partial \hat{\pi}_i}{\partial \theta}\right) S_i^{-1} (y'_i - \hat{\pi}'_i)(y'_i - \hat{\pi}'_i)' S_i^{-1} \left(\frac{\partial \hat{\pi}_i}{\partial \theta}\right)^1$$

Teorema 3.

$\hat{\theta}_G$ merupakan estimator GEE, maka $\lim_{n\to\infty} \sqrt{n}(\hat{\theta}_G - \theta)$ berdistribusi normal multivariat,

$$\lim_{n\to\infty} \sqrt{n}(\hat{\theta}_G - \theta) \sim NM(0; n.Var(\hat{\theta}_G))$$

dengan

$$Var(\hat{\theta}_G) = \left(\sum_{i=1}^{n} W_i \Delta_i S_i^{-1} \Delta_i W_i^{'}\right)^{-1} \left(\sum_{i=1}^{n} W_i \Delta_i S_i^{-1} (Y'_i - \pi'^{(k)}_i)(Y'_i - \pi'^{(k)}_i)' S_i^{-1} \Delta_i W_i\right) \left(\sum_{i=1}^{n} W_i \Delta_i S_i^{-1} \Delta_i W_i^{'}\right)^{-1}$$

Pengujian data simulasi

Selanjutnya akan dilakukan pengujian pada sampel terbatas berdasarkan data simulasi untuk mengamati sifat estimator yang diperoleh dari metode MLE dan metode GEE, serta pengaruh besarnya sampel dan besarnya korelasi. Penelitian dilakukan untuk sampel berukuran n=50, n=100, n=500, n=1000 dan n=5000. Diambil kasus T=3. Model utilitas subjek i memilih alternatif j pada periode t adalah

$U_{ijt} = V_{ijt} + \varepsilon_{ijt}$. untuk t=1,2,3 ; i=1,2,...,n ; j=0,1; $V_{ijt} = \beta_{jt} X_i + \gamma_t Z_{ijt}$, $\beta_{0t}=1$ (16)

Data dibangkitkan pada $\beta_{11}=\beta_1= 0.3$; $\beta_{12}=\beta_2 =0.6$; $\beta_{13}=\beta_3=0.9$, $\gamma_3=1$; $\gamma_3=1.5$; $\gamma_3 = 2$. Nilai variabel observasi X_i dan Z_{ijt} diambil dari distribusi normal, yaitu

$X_i \sim N(0,1)$; $Z_{i0t} \sim N(0,1)$; $Z_{i1t} \sim N(2,1)$

Persamaan (14) ditranformasi ke dalam selisih utilitas

$U_{it} = U_{i0t} - U_{i1t} = (\beta_{1t} - 1)X_i + \gamma_t(Z_{i1t} - Z_{i0t}) = \beta_t X_i + \gamma_t(Z_{i1t} - Z_{i0t})$ (17)

dengan $\beta_1 = -0.7$; $\beta_2 = -0.4$;$\beta_3 = -0.1$; $\gamma_3 = 1$; $\gamma_3 = 1.5$; $\gamma_3 = 2$.

Tiga struktur korelasi, yaitu independen (ρ_0), korelasi sedang (ρ_1), korelasi kuat (ρ_2)

$$\rho_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \rho_1 = \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix},$$

$$\rho_2 = \begin{pmatrix} 1 & 0.9 & 0.9 \\ 0.9 & 1 & 0.9 \\ 0.9 & 0.9 & 1 \end{pmatrix}.$$

Analisis menggunakan software R.2.5.0. khususnya *library (mnormt), library(systemfit), library(micEcon)*, *library(geepack)* yang dapat diakses pada situs www.R-project.org.

HASIL DAN PEMBAHASAN

Hasil estimasi dari perulangan sebanyak 50 kali disajikan dalam Tabel 1. Setiap sel pada tabel 1 memuat rata-rata dan deviasi standar untuk masing-masing parameter. Pada ukuran sampel lebih dari 500, estimator yang dihasilkan sudah cukup baik (variansi dan bias kecil). Semakin besar sampelnya, estimator yang diperoleh semakin mendekati nilai parameter yang sesungguhnya. Pada n=5000, besarnya faktor pengali dari distribusi normal ke distribusi nilai ekstrim adalah ± 1,83. Nugraha (2008) telah menyampaikan bahwa estimator model logit besarnya sekitar 1,6 kali estimator pada model probit.

Pada simulasi berikutnya akan dilakukan pada n=500. Simulasi dilakukan untuk melihat pengaruh tingkat korelasi terhadap estimator. Metode estimasi yang digunakan adalah metode MLE dan metode GEE.

Pada metode MLE digunakan dua jenis asumsi yaitu asumsi korelasi nul/independen ($\theta_{MLE.1}$) dan asumsi korelasi tidak diketahui ($\theta_{MLE.2}$). Pada metode GEE menggunakan dua asumsi, yaitu asumsi independen ($\theta_{GEE.1}$), dan asumsi korelasi tidak diketahui ($\theta_{GEE.2}$). Estimator parameter korelasi dapat dilihat pada $\theta_{MLE.2}$ dan $\theta_{GEE.2}$.

Korelasi rendah (independen).

Dari Tabel 2, estimator $\theta_{MLE.2}$ menghasilkan bias yang lebih besar dibanding dengan $\theta_{MLE.1}$. Walaupun estimator $\theta_{MLE.2}$ tidak stabil, tetapi lebih efisien dibandingkan estimator yang lain. Estimator $\theta_{GEE.1}$ dan $\theta_{GEE.2}$ pada umumnya berbeda, tetapi perbedaanya tidak signifikan, demikian juga nilai efisiensi relatifnya relatif sama. Jadi dapat disimpulkan bahwa pada struktur korelasi rendah, estimator $\theta_{MLE.1}$,$\theta_{GEE.1}$ dan $\theta_{GEE.2}$ adalah sama. Estimasi parameter pada struktur korelasi independen (korelasi rendah), disarankan tidak menggunakan estimator $\theta_{MLE.2}$.

Korelasi sedang (ρ_1)

Dari Tabel 3, estimator $\theta_{MLE.2}$ berbeda secara signifikan terhadap estimator yang lain. Variansi estimator $\theta_{MLE.2}$ dari 50 perulangan adalah cukup besar, yang menggindikasikan estimator ini tidak stabil. Namun demikian estimator $\theta_{MLE.2}$ adalah yang paling efisien. Estimator $\theta_{GEE.2}$ lebih efisien dibandingkan estimator $\theta_{MLE.1}$ dan $\theta_{GEE.1}$. Pada tingkat korelasi menengah disarankan untuk menggunakan estimator $\theta_{MLE.2}$ dan $\theta_{GEE.2}$.

Tabel 1**.** Estimator pada struktur independen dengan n=50, 100, 500, 1000, 5000.

θ	n=50	n=100	n=500	n=1000	n=5000
β_1 =-0.7	-16.72676	-1.382420	-1.3156010	-1.2987803	-1.2882580
	83.72857	0.4595683	0.2147049	0.1409521	0.06273694
β_2 =-0.4	10.57378	-1.029993	-0.6762419	-0.7385833	-0.7466852
	44.74932	0.8257090	0.1809128	0.1716151	0.07369822
β_3 =-0.1	-14.83119	2.453454	-0.1980757	-0.1907568	-0.1831842
	80.43420	18.1880681	0.3155704	0.1536290	0.08231370
γ_1 =1	24.42945	1.940491	1.8597843	1.8322580	1.8198639
	109.50122	0.4588031	0.2052540	0.1038094	0.05386884
γ_2 =1.5	21.31712	3.492314	2.7917158	2.7710875	2.7407762
	127.45044	1.7159711	0.2052540	0.2092163	0.09529880
γ_3 =2	128.62211	29.736070	3.8259477	3.7736695	3.6976281
	259.98991	181.0836355	0.4906767	0.3448178	0.14541870

Tabel 2. Rata-rata selisih estimator dari 50 ulangan korelasi nul menggunakan metode MLE dan GEE.

Parameter θ	$\theta_{MLE.1}$		$\theta_{MLE.2}-\theta_{MLE.1}+\theta$		$\theta_{GEE.1}$		$\theta_{GEE.2}-\theta_{GEE.1}+\theta$	
	mean	sd	mean	sd	mean	sd	mean	sd
β_1 =-1.281	-1.2541	0.2262	-1.2542	0.2262	-1.2787	0.00843	-1.2542	0.2262
β_2 =-0.732	-0.7559	0.2993	-0.7559	0.2993	-0.7339	0.00820	-0.7559	0.2993
β_3 =-0.183	-0.1890	0.2470	-0.1890	0.2470	-0.1841	0.00978	-0.1890	0.2469
γ_1 =1.83	1.7981	0.1520	1.7982	0.1520	1.8282	0.00746	1.7982	0.1519
γ_2 =2.745	2.7770	0.2874	2.7770	0.2874	2.7486	0.01213	2.7771	0.2874
γ_3 =3.66	3.7758	0.4710	3.7758	0.4711	3.6590	0.02514	3.7758	0.4711

Tabel 3. Rata-rata dan standard deviasi estimator pada struktur korelasi ρ_1 metode MLE dan GEE dari 50 kali perulangan

Parameter θ	$\theta_{MLE.1}$		$\theta_{MLE.2}$		$\theta_{GEE.1}$		$\theta_{GEE.2}$	
	mean	sd	mean	sd	mean	sd	mean	sd
β_1 =-1.281	-1.2656	0.2395	-1.2912	0.2015	-1.2869	0.2000	-1.2892	0.3942
β_2 =-0.732	-0.7696	0.2429	-0.7372	0.1794	-0.7399	0.1779	-0.5957	0.3168
β_3 =-0.183	-0.2271	0.2858	-0.2212	0.3140	-0.2211	0.3115	-0.3607	0.2019
γ_1 =1.83	1.7941	0.1947	1.8478	0.1744	1.8430	0.1750	1.9828	0.5958
γ_2 =2.745	2.7811	0.3038	2.8335	0.3032	2.8229	0.2933	2.5538	0.4891
γ_3 =3.66	3.6794	0.4788	3.7155	0.4998	3.7206	0.4956	2.5486	0.7956
ρ_{12}	-	-	0	0	0.1205	0.1205	0.0648	0.0774
ρ_{13}	-	-	0	0	0.0710	0.0698	0.1987	0.2471
ρ_{23}	-	-	0	0	0.0593	0.0533	0.0792	0.0603

Tabel 4. Rata-rata dan standart deviasi estimator pada struktur korelasi ρ_2 metode MLE dari replikasi sebanyak 50 kali

Parameter θ	$\theta_{MLE.1}$		$\theta_{MLE.2}$		$\theta_{GEE.1}$		$\theta_{GEE.2}$	
	mean	sd	mean	sd	mean	sd	mean	sd
β_1 =-1.281	-1.2655	0.2394	-1.3315	0.2406	-1.3323	0.2385	-1.3377	0.3845
β_2 =-0.732	-0.7683	0.2571	-0.7165	0.2030	-0.7134	0.2059	-0.6351	0.3104
β_3 =-0.183	-0.2744	0.3125	-0.1659	0.2617	-0.1588	0.2443	-0.3906	0.2002
γ_1 =1.83	1.7940	0.1947	1.8441	0.1807	1.8417	0.17885	2.0613	0.5326
γ_2 =2.745	2.8571	0.2950	2.8441	0.4008	2.8419	0.4072	2.7965	0.5091
γ_3 =3.66	3.6788	0.4293	3.7843	0.5273	3.7795	0.5038	2.6578	0.6712
ρ_{12}	-	-	0	0	0.2288	0.1298	0.0476	0.0544
ρ_{13}	-	-	0	0	0.1636	0.1186	0.1928	0.2099
ρ_{23}	-	-	0	0	0.1234	0.0967	0.0747	0.0497

Korelasi kuat (ρ_2)

Dari Tabel 4, pada struktur korelasi ρ_2 estimator $\theta_{MLE.2}$ berbeda secara signifikan terhadap estimator lain. Estimator $\theta_{GEE.2}$ lebih efisien dibandingkan dengan estimator $\theta_{GEE.1}$, θ_{MLE}.

Sehingga dapat disimpulkan, jika terdapat korelasi kuat antar respon sebaiknya digunakan estimator $\theta_{GEE.2}$ dan $\theta_{MLE.2}$. Diantara estimator $\theta_{GEE.2}$ dan $\theta_{MLE.2}$, keunggulan masing-masing estimator sangat tergantung pada data yang dianalisis, sehingga tidak dapat digeneralkan.

KESIMPULAN

Pemodelan pada respon biner multivariat dengan variabel independen berupa karakteristik individu dan karakteristik pilihan dapat dilakukan menggunakan model distribusi nilai ekstrim. Metode MLE dan metode GEE dapat digunakan untuk mengestimasi parameter. Kedua metode menghasilkan persamaan penaksir yang harus diselesaikan secara iterasi. Pengujian estimator dapat dilakukan menggunakan pendekatan teori sampel besar.

Pada tingkat korelasi rendah, disarankan tidak menggunakan estimator $\theta_{MLE.2}$. Pada tingkat korelasi sedang maupun korelasi tinggi, sebaiknya digunakan estimator $\theta_{MLE.2}$ dan $\theta_{GEE.2}$. Estimator parameter korelasi pada metode $\theta_{MLE.2}$ dan $\theta_{GEE.3}$ tidak dapat digunakan untuk menggambarkan tingkat korelasi yang sesungguhnya.

DAFTAR PUSTAKA

Contoyannis P, Andrew MJ, Gonzales RL. 2002. *Using Simulation-Based Inference With Panel Data In Health Economics*. Working Paper Department of Economics and Related Studies, University of York.

Fitzmaurice GM, Laird NM, Ratnitzky AG. 1993. Regression Models for DiscreteLongitudinal Responses. *Statistical Science* Vol. **8** No. 3 : 284 – 309

Harris MN, Macquarie LR, Siouclis AJ. 2000, Comparison of alternative Estimators for Binary Panel Probit Models, *Melbourne Institute Working Paper* no 3/00

Horowitz JL & Savin NE. 2001. Binary Response Models : Logits, Probits and Semiparametrics, *Journal of Economic Perspectives,* Volume **15(**4): 43-56

Nugraha J, Haryatmi S, Guritno S. 2006. Model Discrete Choice dan Regresi Logistik, *Proseding Seminar Nasional MIPA*, UNY.

Nugraha J.2008. Model Probit dan Model Logit pada Respon Biner. *Eksakta*, *Jurnal Ilmu-ilmu MIPA* Vol **10**(1): 9-17

Liang KY & Zeger SL. 1986. Longitudinal Data Analysis Using Generalised Linear Models, *Biometrika* **73**: 13-22.

Prentice.1988. Correlated Binary Regression with Covariates Specific to Each Binary Observation. *Biometrics* **44** : 1043-1048.

R Development Core Team (2008). *R: A language and Environment for Statistical Computing. R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org, [24-5-2008, 2:13]

Train K .2003. *Discrete Choice Methods with Simulation*, UK Press, Cambridge.