Model Regresi Logistik untuk Respons Biner Multivariate dengan Generalized Estimating Equation

Oleh :
Jaka Nugraha[1], Suryo Guritno[2] & Sri Haryatmi Kartiko[2]
1. Jurusan Statistika, FMIPA UII. Email: jnugraha@fmipa.uii.ac.id
2.Jurusan Matematika FMIPA UGM

Abstrak

Dalam makalah ini akan dibahas model regresi untuk data kategorik multivariate, khususnya respon biner (dikotomis). Untuk mengestimasi parameter pada model marginal, akan digunakan pendekatan *Independent Estimating Equation* (IEE) dan pendekatan *Generalized Estimating Equations* (GEE). Proses iterasi menggunakan metode scoring untuk menyelesaikan persamaan penaksir dengan program R. Dari hasil simulasi, diperoleh bahwa pendekatan GEE menghasilkan penaksir parameter dengan variansi yang lebih kecil dibandingkan dengan pendekatan IEE.
Kata kunci : model marginal, fungsi link,IEE, GEE

Regression Logistic Model for Multivariate biner Response by Generalized Estimating Equation

Abstract

This paper will discuss about regression model for multivariate categorical response, especially for biner response (dichotomus). To estimate paramters in marginal model, it will used approximation Independent Estimating Equation (IEE) and Generalized Estimating Equations (GEE). Iteration process performade by scoring methods to solve estimating equation using R programe. Results show that GEE approach give lower variance estimating parameter than that of IEE approach
Key words : marginal model, link functions, IEE, GEE

1. Pengantar

Analisis pada respon data kontinu dan diasumsikan berdistribusi Gausian, biasanya digunakan analisis model linear. Dalam respon biner, yang biasanya dipilih model non linear untuk mean, model ini akan membuat kesulitan interpretasi model marginalnya. Model efek random, sejauh ini sangat sedikit yang dikembangkan untuk respon biner dibandingkan dengan respon kontinu. GEE merupakan pendekatan yang menarik untuk respon biner multivariate, yang mana tidak memerlukan spesifikasi lengkap pada distribusi gabungan.

Diasumsikan bahwa N individu masing-masing diobservasi sebanyak J respon. Y_{ij} adalah respon ke j pada individu/subjek ke i dan setiap responnya adalah biner. Sehingga respon pada individu ke i, dapat disajikan dalam bentuk $Y_i = (Y_{i1},....,Y_{iJ})$ sebagai vektor 1xJ, dimana variabel random biner $Y_{ij} = 1$ jika subjek i mempunyai nilai 1 (sukses) pada respon j=1 dan 0 untuk yang lain. Setiap individu mempunyai vektor covariate x_{ij} berukuran Px1 untuk setiap respon j dan $X_i = (x_{i1},...,x_{iJ})^t$ merupakan matrik kovariate JxP untuk individu i. Sehingga data untuk individu ke-i berisi observasi (Y_i,X_i).

Dipresentasikan dalam Seminar Nasional MIPA 2006 dengan tema **"Penelitian, Pendidikan, dan Penerapan MIPA serta Peranannya dalam Peningkatan Keprofesionalan Pendidik dan Tenaga Kependidikan"** yang diselenggarakan oleh Fakultas MIPA UNY, Yogyakarta pada tanggal 1 Agustus 2006

Distribusi marginal Y_{ij} adalah Bernouli dan model marginalnya adalah [6]

$f(y_{ij}/X_i) = \exp[y_{ij}\theta_{ij} - \log\{1+\exp(\theta_{ij})\}]$,

(1)

dimana diasumsikan bahwa $\theta_{ij} = \log[\pi_{ij}/(1-\pi_{ij})] = X_{ij}^t\beta$ dan $\pi_{ij} = \pi_{ij}(\beta) = E(Y_{ij}) = pr(Y_{ij}=1|x_{ij}, \beta_j)$ yang merupakan probabilitas sukses untuk respon j dan β_j adalah vektor parameter berukuran Px1. Dalam hal ini kita hanya membuat asumsi pada distribusi marginal Y_{ij}. Model marginal dapat dituliskan sebagai

$$\pi_{ij} = pr(Y_{ij}=1|x_{ij}, \beta) = \frac{\exp(X_{ij}^T\beta_j)}{1+\exp(X_{ij}^T\beta_j)} \text{ untuk } j=1,2,\ldots J$$

(2)

dan $X_{ij} = (X_{1ij}, \ldots.., X_{pij})^T$ adalah matrik px1. Model ini dinamakan model marginal regresi logistik. Selanjutnya $\pi_{ij}(\beta)$ dapat dikelompokkan bersama-sama ke bentuk vektor $\pi_i(\beta)$ yang memuat probabilitas marginal sukses, $\pi_i(\beta) = E(Y_i) = (\pi_{i1}, ..., \pi_{iJ})^t$. Pada respon biner, fungsi link logit adalah yang biasa dipilih, pada prinsipnya sebarang fungsi link dapat digunakan.

Dalam makalah ini akan dibahas dua pendekatan estimasi parameter, yaitu IEE dan GEE dari data simulasi dengan program R. Diambil kasus untuk data biner bivariate dan tri variate. Sebagai pembanding hasil, data juga diestimasi secara parsial dengan *generalized linear model.*

2. Model Observasi Independen

Jika diantara J respon diasumsikan saling independen, maka distribusi bersama respons biner tersebut adalah

$$f(y_i|X_i) = \exp\left[\sum_{j=1}^{J} y_{ij}\theta_{ij} - \sum_{j=1}^{J}\log[1+\exp(\theta_{ij})]\right]$$

(3)

Penaksiran parameter regresi logistik dengan pendekatan IEE mengasumsikan bahwa vektor $Y_1,\ldots,Y_n$ sebagai ulangan observasi adalah independen dan pasangan observasi marginal $Y_{i1},\ldots Y_{iJ}$ juga saling independent independen.

Teorema 1.

Penaksir untuk IEE dari β katakanlah $\hat{\beta}_I$ adalah suatu penyelesaian dari persamaan

$\sum_{i=1}^{n} X_i(Y_i^T - \pi_i^T) = 0$ dengan $X_i = (X_{i1},\ldots. X_{iT})$ dan $\pi_i = (\pi_{i1}, \ldots,\pi_{iT})$, Y_i adalah vektor observasi.

Bukti:

Y_{ij} dengan j=1,2,..J dan i=1, ….,n merupakan kejadian Bernouli, sehingga densitas marginal Y_{ij} dapat dinyatakan sebagai

$$f(Y_{ij}) = \exp[\, y_{ij}\theta_{ij} - \ln\{1 + \exp(\theta_{ij})\}]$$

$$\theta_i = \begin{bmatrix} \theta_{i1} \\ \theta_{i2} \\ \dots \\ \theta_{iT} \end{bmatrix}$$

Karena $Y_1,\ldots,Y_n$ sebagai ulangan observasi adalah independen, demikian juga pasangan observasi marginal $Y_{i1},\ldots Y_{iT}$ independen maka fungsi log-kemungkinan untuk individu ke-i adalah

$$f(Y_i) = \exp[\sum_j y_{ij}\theta_{ij} - \sum_j \ln[1 + \exp(\theta_{ij})]$$

$$L_i = \ln f(Y_i) = \sum_j y_{ij}\theta_{ij} - \sum_j \ln[1 + \exp(\theta_{ij})]$$

Kemudian fungsi di atas diderivatifkan terhadap β untuk individu ke-i adalah

$$\frac{\partial L_i}{\partial \beta} = \frac{\partial \pi_i}{\partial \beta}\frac{\partial \theta_i}{\partial \pi_i}\frac{\partial L_i}{\partial \theta_i}$$

selanjutnya

1. $$\frac{\partial L_i}{\partial \theta_{ij}} = y_{ij} - \frac{\exp(\theta_{ij})}{1 + \exp(\theta_{ij})}$$

$$= y_{ij} - E(y_{ij})$$

$$= y_{ij} - \pi_{ij}$$

$$\frac{\partial L_i}{\partial \theta_i} = \begin{bmatrix} \frac{\partial L_i}{\partial \theta_{i1}} \\ \frac{\partial L_i}{\partial \theta_{i2}} \\ \dots\dots \\ \frac{\partial L_i}{\partial \theta_{iT}}. \end{bmatrix} = \begin{bmatrix} y_{i1} - \pi_{i1} \\ y_{i2} - \pi_{i2} \\ \dots\dots\dots \\ y_{iT} - \pi_{iT} \end{bmatrix} = Y_i^T - \pi_i^T$$

2. $$\frac{\partial \pi_i}{\partial \theta_i} = \begin{bmatrix} \frac{\partial \pi_{i1}}{\partial \theta_{i1}} & \frac{\partial \pi_{i2}}{\partial \theta_{i1}} & \cdots\cdots & \frac{\partial \pi_{iT}}{\partial \theta_{i1}} \\ \frac{\partial \pi_{i1}}{\partial \theta_{i2}} & \frac{\partial \pi_{i2}}{\partial \theta_{i2}} & \cdots\cdots & \frac{\partial \pi_{iT}}{\partial \theta_{i2}} \\ \cdots\cdots & & & \\ \frac{\partial \pi_{i1}}{\partial \theta_{iT}} & \frac{\partial \pi_{i2}}{\partial \theta_{iT}} & \cdots\cdots & \frac{\partial \pi_{iT}}{\partial \theta_{iT}} \end{bmatrix}$$

$$\frac{\partial \pi_{it}}{\partial \theta_{ij}} = \begin{cases} \frac{\exp(\theta_{ij})}{[1+\exp(\theta_{ij})]^2} = \pi_{ij}(1-\pi_{ij}) = \mathrm{Var}(Y_{ij}) & \text{jika } j = t \\ 0 & \text{jika } j \neq t \end{cases}$$

$$\text{jadi } \frac{\partial \pi_i}{\partial \theta_i} = \begin{bmatrix} \mathrm{Var}(Y_{i1}) & 0 & \ldots & 0 \\ 0 & \mathrm{Var}(Y_{i2}) & \ldots & 0 \\ \ldots & \ldots & \ldots. & \ldots \\ 0 & 0 & \ldots. & \mathrm{Var}(Y_{iT}) \end{bmatrix} = \mathrm{Diag}\{\ \mathrm{Cov}(Y_i)\}$$

3. $$\frac{\partial \pi_i}{\partial \beta} = \begin{pmatrix} \frac{\partial \pi_{i1}}{\partial \beta} & \frac{\partial \pi_{i2}}{\partial \beta} & \ldots. & \frac{\partial \pi_{iT}}{\partial \beta} \end{pmatrix}$$

Karena $\pi_{ij} = \dfrac{\exp(X_{ij}^T \beta_j)}{1+\exp(X_{ij}^T \beta_j)}$, maka

$$\frac{\partial \pi_{ij}}{\partial \beta_j} = X_{ij} \cdot \frac{\exp(X_{ij}^T \beta_j)}{[1+\exp(X_{ij}^T \beta_j)]^2} = X_{ij}\mathrm{Var}(Y_{ij})$$

Sehingga

$$\frac{\partial \pi_i}{\partial \beta} = \begin{bmatrix} X_{i1}Var(Y_{i1}) & X_{i2}Var(Y_{i2}) & \ldots. & X_{iT}Var(Y_{iT}) \end{bmatrix}$$

$$\frac{\partial \pi_i}{\partial \beta} = \begin{bmatrix} x_{i11}\mathrm{Var}(Y_{i1}) & x_{i21}\mathrm{Var}(Y_{i2}) & \ldots & x_{iT1}\mathrm{Var}(Y_{iT}) \\ x_{i12}\mathrm{Var}(Y_{i1}) & x_{i22}\mathrm{Var}(Y_{i2}) & \ldots & x_{iT2}\mathrm{Var}(Y_{iT}) \\ \ldots & \ldots & \ldots & \ldots. \\ x_{i1p}\mathrm{Var}(Y_{i1}) & x_{i2p}\mathrm{Var}(Y_{i2}) & \ldots. & x_{iTp}\mathrm{Var}(Y_{iT}) \end{bmatrix} = X_i \Delta_i$$

dengan $\Delta_i = \mathrm{Diag}\{\mathrm{Cov}(Y_i)\}$

Dari (1), (2) dan (3) diperoleh

$$\frac{\partial L_i}{\partial \beta} = \frac{\partial \pi_i}{\partial \beta}\frac{\partial \theta_i}{\partial \pi_i}\frac{\partial L_i}{\partial \theta_i} = X_i \mathrm{Diag}\{\mathrm{Cov}(Y_i)\}[\mathrm{Diag}\{\ \mathrm{Cov}(Y_i)\}]^{-1}\ (Y_i^T - \pi_i^T)$$

$$0 = X_i\ (Y_i^T - \pi_i^T)$$

Kemudian penaksir kemungkinan maksimum untuk parameter β adalah penyelesaian dari

$$\sum_{i=1}^{n}\frac{\partial L_i}{\partial \beta}=\sum_{i=1}^{n}\frac{\partial \pi_i}{\partial \beta}\frac{\partial \theta_i}{\partial \pi_i}\frac{\partial L_i}{\partial \theta_i}=\sum_{i=1}^{n}X_i\,(y_i^T-\pi_i^T)=0 \qquad \blacklozenge$$

Derevatif ke dua dari fungsi likelihood adalah

$$\frac{\partial}{\partial \beta^T}\left(\frac{\partial L_i}{\partial \beta}\right)=\frac{\partial}{\partial \beta^T}\left(X_i\,(y_i^T-\pi_i^T)\right)$$

$$=-X_i\,\frac{\partial \pi_i^T}{\partial \beta^T}=-X_i\,(X_i\,\Delta_i)^T$$

$$=-X_i\,\Delta_i X_i^T=-X_i\,[diag(Cov(Y_i)]X_i^T$$

$I(\beta)=-\sum_{i=1}^{n}X_i\,\Delta_i X_i^T$ biasa disebut dengan matrik informasi.

Lemma 1.

Penyelesaian persamaan penaksir IEE dengan menggunakan metode Newton-Raphson pada iterasi ke-t+1 adalah

$$\beta^{(t+1)}=\beta^{(t)}+\left(\sum_{i=1}^{n}X_i\mathrm{Diag}\{\pi_i^{(t)}(1-\pi_i^{(t)})\}X_i^T\right)^{-1}\left(\sum_{i=1}^{n}X_i(Y_i^T-\pi_i^{T(t)})\right) \qquad (4)$$

Bukti:

Rumus dari metode newton rampson adalah

$$\beta^{(t+1)}=\beta^{(t)}-(H^{(t)})^{-1}q^{(t)}$$

diasumsikan $H^{(t)}$ nonsingular.

$$q^T=\left[\frac{\partial L(\beta)}{\partial \beta_1}\quad\ldots\quad\frac{\partial L(\beta)}{\partial \beta_p}\right]\text{ dan }H=\begin{bmatrix}\frac{\partial^2 L(\beta)}{\partial \beta_1^2} & \ldots & \frac{\partial^2 L(\beta)}{\partial \beta_p \partial \beta_1}\\ \ldots & \ldots & \ldots\\ \frac{\partial^2 L(\beta)}{\partial \beta_p \partial \beta_1} & \ldots & \frac{\partial^2 L(\beta)}{\partial \beta_p^2}\end{bmatrix}$$

$q^{(t)}$ dan $H^{(t)}$ merupakan suku taksiran pada $\beta^{(t)}$, yaitu penduga ke-t dari $\hat{\beta}$ (t = 0,1,2,…).

Di atas telah dihitung bahwa

$$q^{(t)} = \sum_{i=1}^{n} \frac{\partial L_i}{\partial \beta} = \sum_{i=1}^{n} X_i (y_i^T - \pi_i^T)$$

dan

$$H = I(\beta) = -\sum_{i=1}^{n} X_i \Delta_i X_i^T$$

$$= -\sum_{i=1}^{n} X_i \, Diag[\pi_i(1-\pi_i)] X_i^T$$

$$H^{(t)} = -\sum_{i=1}^{n} X_i \mathrm{Diag}\{\pi_i^{(t)}(1-\pi_i^{(t)})\} X_i^T$$

Sehingga terbukti bahwa

$$\beta^{(t+1)} = \beta^{(t)} + \left(\sum_{i=1}^{n} X_i \mathrm{Diag}\{\pi_i^{(t)}(1-\pi_i^{(t)})\} X_i^T\right)^{-1} \left(\sum_{i=1}^{n} X_i (Y_i^T - \pi_i^{T(t)})\right) \quad \blacklozenge$$

Selain metode Newton Rapson, metode lain yang kadang-kadang lebih sederhana adalah metode scoring [1]. Metode scoring diperoleh dengan mengganti matrik H, yaitu derevatif kedua dari persamaan, diganti dengan matrik E(H).

$$E(H) = E\left[\frac{\partial^2 L(\beta)}{\partial \beta_a \partial \beta_b}\right] = -X\mathrm{Diag}[\hat{\pi}_i (1-\hat{\pi}_i)] X^T$$

Sehingga persamaan iterasi dengan metode scoring adalah

$$\beta^{(t+1)} = \beta^{(t)} + \left(\sum_{i=1}^{n} X_i \mathrm{Diag}\{\hat{\pi}_i^{(t)}(1-\hat{\pi}_i^{(t)})\} X_i^T\right)^{-1} \left(\sum_{i=1}^{n} X_i (Y_i^T - \pi_i^{T(t)})\right)$$

(5)

Matrik kovariansinya adalah

$$\mathrm{Cov}(\hat{\beta}) = \left(\sum_{i=1}^{n} X_i \mathrm{Diag}\{\hat{\pi}_i (1-\hat{\pi}_i)\} X_i^T\right)^{-1} \left(\sum_{i=1}^{n} X_i \mathrm{Cov}(Y_i^T) X_i^T\right) \left(\sum_{i=1}^{n} X_i \mathrm{Diag}\{\hat{\pi}_i (1-\hat{\pi}_i)\} X_i^T\right)^{-1}$$

(6)

Penaksir kemungkinan maksimum dengan pendekatan IEE di atas menghasilkan penaksir yang selaras dan Normal asimtotis [2].

Teorema 2.

$\hat{\beta}_I$ adalah penaksir untuk β yang selaras dan $n^{1/2}(\hat{\beta}_I-\beta)$ berdistribusi asimtotis Multivariate Normal (Gausian) pada $n \to \infty$ dengan mean nol dan matrik kovariannya

$$\lim_{n\to\infty} n\left(\sum_{i=1}^{n} X_i Cov(Y_i^T) X_i^T\right)^{-1}$$

Bukti:

Andaikan b sebagai penaksir dari β berdasarkan pendekatan deret Taylor orde pertama, vektor skore $U_{(\beta)}$ untuk nilai β = b adalah

$$U_{(\beta)} \cong U(b) + H(b)(\beta - b)$$

dengan H(b) = I(b), yaitu matrik derivatif ke-dua dari fungsi log-kemungkinan pada β=b dan UU^T=-H. Diketahui bahwa

$$\gamma = E(-H) = \left(\sum_{i=1}^{n} X_i Cov(Y_i^T) X_i^T\right)$$

Selanjutnya untuk sampel besar $U_{(\beta)} \cong U(b) - \gamma(\beta - b)$

Karena U(b) = 0 maka $(b - \beta) \cong \gamma^{-1} U_{(\beta)}$

Diasumsikan γ konstanta dan non singular, dan karena $E\{U_{(\beta)}\} = 0$ maka

$$E(b - \beta) \cong \gamma^{-1} E\{U_{(\beta)}\} = 0$$

sehingga b adalah penaksir tak bias asimtotis untuk β.

Matrik kovariansi untuk b adalah

$$E[(b - \beta)(b - \beta)^T] \cong \gamma^{-1} E(UU^T) \gamma^{-1} = \gamma^{-1}$$

sebab $\gamma = E(UU^T)$ dan γ^{-1} matrik simetris.

Untuk sampel besar

$$(b - \beta) \sim N(\mathbf{0}, \gamma^{-1})$$

$$n^{1/2} (b - \beta) \sim N(\mathbf{0}, n\gamma^{-1})$$ ♦

Pada umumnya, perulangan pada individu yang sama mengakibatkan adanya korelasi, hal ini berakibat invers dari matrik informasi tidak selaras [3].

Dengan adanya asumsi independen antar respon, MLE pada regrei logistik menghasilkan estimasi yang konsisten dan Asymtotis normal [2]. Namun secara umum, jika terdapat korelasi antar respon biner, mengakibatkan invers dari estimator matrik informasi menjadi tidak konsisten. Liang dan Zeger [3] mengusulkan penggunaan estimator "robust" untuk menaksir variansi yang konsisten terhadap korelasi antar respon. Ketika korelasi antar

respon tidak terlalu besar, Lipsitz, S.R., Laird, N.M., dan Harrington, D.P. [5] mengusulkan estimator MLE di atas masih cukup efisien.

3. Pendekatan GEE

Untuk meningkatkan efisiensi penaksiran parameter model marginal, Liang dan Zeger [3,4] dan Prentice [7] telah mengembangkan GEE. Pendekatan GEE menghasilkan estimator konsisten untuk parameter regresi, di bawah specifikasi yang benar untuk fungsi mean, π_i yang merupakan vektor respons untuk masing-masing individu. GEE untuk β didefinisikan sebagai

$$U(\beta) = \sum_{i=1}^{n} D_i^T V_i^{-1} (Y_i^T - \pi_i^T) = 0 \quad (7)$$

dengan $D_i = \frac{\partial \pi}{\partial \beta} = \Delta_i X_i^T$ dan V_i adalah pendekatan untuk matrik kovariansi. V_i dapat dituliskan sebagai

$$V_i = \Delta_i^{1/2} R_i(\alpha) \Delta_i^{1/2}$$

dengan $\Delta_i = \text{Diag}\{\text{Cov}(Y_i)\}$ dan didefinisikan $\Delta_i^{1/2} = \text{Diag}\{\sqrt{\text{var}(y_{i1})} \ \ldots \sqrt{\text{var}(y_{iT})}\}$

$R_i(\alpha) = \text{Corr}(Y_i)$ merupakan matrik JxJ. α menunjukkan vektor parameter yang berkaitan dengan model tertentu untuk $\text{Corr}(Y_i)$.

Koefisien korelasi berpasangan diasumsikan $\rho_{isr}(\alpha) = \text{corr}(Y_{is}, Y_{ir})$ untuk i = 1,2,..n dan s,r=1,2,…J . Untuk mengestimasi R_i, didefinisikan J(J-1)/2 vektor korelasi empirik, r_i dengan elemen-elemen

$$r_{ist} = \frac{(Y_{is} - \pi_{is})(Y_{it} - \pi_{it})}{[\pi_{is}(1-\pi_{is})\pi_{it}(1-\pi_{it})]^{1/2}} \quad (8)$$

Catatan bahwa $E(r_{irs}) = \rho_{irs} = \text{corr}(Y_{is}, Y_{it})$.

GEE untuk regresi logistik dengan menggunakan matrik korelasi $R_i(\alpha)$ adalah

$$U(\beta) = \sum_{i=1}^{n} X_i \Delta_i V_i^{-1} (Y_i - \pi_i) = 0 \quad (9)$$

Persamaan ini dapat diselesaikan dengan metode Newton-Raphson ataupun dengan metode Scoring. Persamaan iterasinya adalah

$$\beta^{(t+1)} = \beta^{(t)} + \left(\sum_{i=1}^{n} X_i \Delta_i \, V_i^{-1} \Delta_i X_i^T \right)^{-1} \left(\sum_{i=1}^{n} X_i \Delta_i \, V_i^{-1} (Y_i^T - \pi_i^{(t)T} \right)$$

(10)

Liang dan Zeger (1989)[4], menunjukkan $\hat{\beta}_{GEE}$ adalah selaras dan berdistribusi normal secara asimtotis dengan matrik kovarian

$$\text{cov}(\hat{\beta}_{GEE}) = \lim_{n\to\infty} \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i\right)^{-1} \left(\sum_{i=1}^{n} D_i^T V_i^{-1} \text{Cov}(Y_i) V_i^{-1} D_i\right) \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i\right)^{-1}$$

(11)

atau dapat dituliskan sebagai

$$\text{Cov}(\hat{\beta}) = \left(\sum_{i=1}^{n} X_i \Delta_i \, V_i^{-1} \Delta_i X_i^T \right)^{-1} \left(\sum_{i} X_i \Delta_i V_i^{-1} \text{Cov}(Y_i^T) V_i^{-1} \Delta_i X_i^T \right) \left(\sum_{i=1}^{n} X_i \Delta_i \, V_i^{-1} \Delta_i X_i^T \right)^{-1}$$

(12)

Penaksiran cov($\hat{\beta}_{GEE}$) dapat diperoleh dengan mengganti Cov(Y_i) dengan $(Y_i - \hat{p}_i)(Y_i - \hat{p}_i)^T$ dan mengganti parameter ρ dengan $\hat{\rho}$. Penaksir cov($\hat{\beta}_{GEE}$) *ini "robust"* karena selaras meskipun $V_i \neq \text{Cov}(Y_i)$.

Jika pemilihan R(α) tepat, dalam arti R(α) menyatakan korelasi sesungguhnya dari Y_i maka $V_i = \text{Cov}(Y_i)$. $\hat{\beta}_{opt}$ adalah penaksir "optimal" untuk β yang merupakan penyelesaian dari GEE pada $V_i = \text{Cov}(Y_i)$, sehingga [3]

$$\text{cov}(\hat{\beta}_{opt}) = \lim_{n\to\infty} \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i\right)^{-1} \left(\sum_{i=1}^{n} D_i^T V_i^{-1} V_i V_i^{-1} D_i\right) \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i\right)^{-1}$$

$$= \lim_{n\to\infty} \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i\right)^{-1}$$

(13)

Misal untuk respon bivariate (J=2), dan korelasi berpasangan ditaksir dengan [5]

$$\rho_i = r_{i12} = \frac{(Y_{i1} - \pi_{i1})(Y_{i2} - \pi_{i2})}{[\pi_{i1}(1-\pi_{i1})\pi_{i2}(1-\pi_{i2})]^{1/2}}$$

(14)

adalah korelasi individu ke- i. dan

$$V_i = \begin{pmatrix} p_{i1}q_{i1} & \rho_i \sqrt{p_{i1}q_{i1}} \sqrt{p_{i2}q_{i2}} \\ \rho_i \sqrt{p_{i1}q_{i1}} \sqrt{p_{i2}q_{i2}} & p_{i2}q_{i2} \end{pmatrix}$$

(15)

Untuk respon trivariate (J= 3),

$$V_i = \begin{pmatrix} p_{i1}q_{i1} & \rho_{i12}\sqrt{p_{i1}q_{i1}p_{i2}q_{i2}} & \rho_{i13}\sqrt{p_{i1}q_{i1}p_{i3}q_{i3}} \\ \rho_{i12}\sqrt{p_{i2}q_{i2}p_{i1}q_{i1}} & p_{i2}q_{i2} & \rho_{i23}\sqrt{p_{i2}q_{i2}p_{i3}q_{i3}} \\ \rho_{i13}\sqrt{p_{i1}q_{i1}p_{i3}q_{i3}} & \rho_{i23}\sqrt{p_{i2}q_{i2}p_{i3}q_{i3}} & p_{i3}q_{i3} \end{pmatrix}$$

(16)

4. Simulasi Estimasi parameter dengan IEE dan GEE

Dalam bab ini akan dilakukan estimasi parameter menggunakan GEE dan IEE berdasarkan data yang dibangkitkan dengan variabel independen dan probabilitas sukses diketahui. Dalam hal ini diambil kasus untuk data bivariat dan trivariate dengan satu variabel indepenen. Fungsi linearnya adalah

$f(x_i) = a + bx_i$

dengan a=1 dan b=-1. Probabilitas sukses (π_i) adalah

$\pi_i = \exp\{f(x_i)\}/[1+ \exp\{f(x_i)\}]$

Proses perhitungan menggunakan program R 2.30. dengan komputer RAM 256 Prosesor Intel Pentium 1,76 Giga. Simulasi dilakukan untuk beberapa n, yaitu 100, 500, 1000 dan 2000. Langkah-langkah simulasi adalah sbb:

1. bangkitkan data untuk variable independent
   $x1 \sim N(\mu,\sigma^2)$ dan $z1 \sim UNIF(0,1)$, $z2 \sim UNIF(0,1)$
   $x2= x1 + z1$ dan $x3 = x1 + z2$
2. hitung probabilitas π_{i1}, π_{i2}, π_{i3} berdasarkan variabel independen tersebut
3. Dari nilai probabilitas yang terbentuk, bangkitkan data Y_{ij} = binomial (π_{ij}). j=1,2,3 dan i= 1,2, …n
4. Cari penaksir parameter dengan metode Newton Raphson dan scoring untuk cara IIE dan GEE. Sebagai pembanding, dilakukan estimasi parameter dengan GLM secara parsial
5. Hitung efisiensi antara IEE dan GEE untuk masing-masing penaksir robust dan penaksir optimal.

Dari proses simulasi yang telah dilakukan, dapat dicatat beberapa hal yaitu

1. dari rumus korelasi r_i, berarti harus disyaratkan bahwa $\pi_{ij} \neq 0$ ataupun $\pi_{ij} \neq 1$.
2. untuk GEE disyaratkan bahwa matrik informasi harus non singular atau matrik V_i adalah non singular.

Hasil simulasi adalah

1. Secara umum GEE menghasilkan penaksir yang mempunyai variansi lebih kecil dibandingkan dengan IEE, baik untuk penasir robust maupun penaksir optimal

2. Penaksir untuk IEE nilainya lebih dekat dengan nilai parameter yang sesungguhnya (a=1 dan b= -1) dibanding penaksir GEE.
3. Penaksir optimal untuk IEE sama dengan penaksir GLM yang dilakukan secara parsial (masing masing Y_j secara terpisah).

5. Simpulan

. Pendekatan GEE menarik, yang dapat diringkas sebagai berikut

1. memberikan penaksir yang konsisten untuk $\hat{\beta}$, dan hanya memerlukan spesifikasi untuk $\pi_i(\beta)$ dan menghasilkan penaksir dengan variansi yang lebih kecil dibanding penaksir IEE.
2. Jika pemilihan korelasi dilakukan secara tepat, maka akan diperoleh penaksir yang konsisten.
3. walaupun pemilihan corelation tidak diperoleh secara tepat, penaksir yang konsisten masih dapat diperoleh dengan penaksir *"robust"*.
4. kelemahan utama pendekatan GEE adalah dalam proses iterasi diperoleh matrikV_i yang singular.

Daftar Pustaka

[1] Agresti A. (1990), **Categorical Data Analysis**, John Wiley & Son

[2] Fitzmaurice, G.M., Laird N.M., Ratnitzky, A.G. (1993) Regression Models for Discrete Longitudinal Responses. ***Statistical Science*** Vol. 8 No. 3. 284 – 309

[3] Liang, K.Y., dan Zeger,S.L (1986). Longitudinal data analysis using generalised linear models, ***Biometrika*** 73, 13-22.

[4] Liang, K.Y., dan Zeger, S.L., (1989). A class of logistic regression models for multivariate binary time series. ***Journal of the American Statistical Assosiation*** 84, 447-457.

[5] Lipsitz, S.R., Laird, N.M., dan Harrington, D.P. (1990). Maximum likelihood regression models for paired binary data. ***Statistics in Medicine*** 9, 1517-1525.

[6] McDonald B.W. (1993) Estimating Logistic Regrssion Parameters for Bivariate Binary Data. ***Journal of The Royal Statistical Society,*** Series B 55, 391 - 397.

[7] Prentice, R.L. (1988) Correlated binary regression with covariates specific to each binary observation. ***Biometrics*** 44, 1033-1048.

LAMPIRAN

Tabel 1. Output program estimasi parameter data bivariate pada N=100

GLM	BETA	SE	Z Value		
A1	5.045	2.948	1.711		
B1	-2.265	1.135	-1.996		
A2	0.9563	0.5344	1.789		
B2	-0.6504	-0.1636	-3.976		
IEE ,R[i]=0	**BETA**	**Cov1**	**Z Value**	**Cov2**	**Z value**
A1	5.0448555	14.22186678	0.3547253	2.9483166	1.711097
B1	-2.2650045	5.35821544	-0.4227162	1.1348511	-1.995861
A2	0.9562733	0.45173530	2.1168886	0.5344681	1.789206
B2	-0.6503617	0.08573539	-7.5856860	0.1636298	-3.974592
GEE ,R[i]	**BETA**	**Cov1**	**Z Value**	**Cov2**	**Z value**
A1	1.8015049	0.29432632	6.120774	0.40142466	4.487778
B1	-0.3165046	0.03838640	-8.245226	0.06254388	-5.060520
A2	1.5186398	0.24241165	6.264715	0.35549377	4.271917
B2	-0.2849686	0.02837773	-10.041979	0.05431604	-5.246490

Tabel 2. Output program estimasi parameter data bivariate pada N=500

GLM	BETA	SE	Z Value		
A1	1.0701	0.2692	3.975		
B1	-1.1592	0.1622	-7.148		
A2	0.8781	0.2405	3.652		
B2	-0.9912	0.1256	-7.891		
IEE ,R[i]=0	**BETA**	**Cov1**	**Z Value**	**Cov2**	**Z value**
A1	1.0700872	0.2979654	3.591314	0.2691859	3.975272
B1	-1.1591970	0.2001041	-5.792970	0.1621758	-7.147779
A2	0.8781094	0.2330070	3.768596	0.2404563	3.651846
B2	-0.9912273	0.1237122	-8.012365	0.1256124	-7.891159
GEE ,R[i]	**BETA**	**Cov1**	**Z Value**	**Cov2**	**Z value**
A1	1.0031520	0.05821698	17.23126	0.08931538	11.231571
B1	-0.5770756	0.03057907	-18.87159	0.05827349	-9.902883
A2	1.0479571	0.06019969	17.40801	0.09209661	11.378889
B2	-0.5760072	0.03092896	-18.62356	0.05824804	-9.888870

Tabel 3. Output program estimasi parameter data bivariate pada N=1000

GLM	BETA	SE	Z Value		
A1	0.76996	0.17653	4.362		
B1	-1.00961	0.09367	-10.778		
A2	0.92402	0.18058	5.117		
B2	-0.94696	0.08527	-11.105		
IEE ,R[i]=0	**BETA**	**Cov1**	**Z Value**	**Cov2**	**Z value**
A1	0.7699600	0.16899768	4.556039	0.17652913	4.361660
B1	-1.0096104	0.09606828	-10.509301	0.09366973	-10.778407
A2	0.9240239	0.17330637	5.331736	0.18057944	5.116994
B2	-0.9469636	0.07899588	-11.987506	0.08527413	-11.104934
GEE ,R[i]	**BETA**	**Cov1**	**Z Value**	**Cov2**	**Z value**
A1	0.6818764	0.12184063	5.596462	0.14230838	4.791541
B1	-0.7603083	0.04254192	-17.871979	0.06142647	-12.377535
A2	0.8870075	0.13214483	6.712389	0.15093139	5.876892
B2	-0.7584685	0.04251865	-17.838489	0.06123481	-12.386230

Tabel 4. Output program estimasi parameter data Trivariate pada N=100

GLM	**BETA**	**SE**	**Z Value**		
A1	1.0544	0.2587	4.075		
B1	-1.0280	0.2735	-3.759		
A2	1.2112	0.3286	3.686		
B2	-1.5961	0.3533	-4.517		
A3	0.8078	0.2732	2.956		
B3	-0.9304	0.2348	-3.962		
IEE ,R[i]=0	**BETA**	**Cov1**	**Z Value**	**Cov2**	**Z value**
A1	1.0543550	0.2638117	3.996620	0.2587494	4.074812
B1	-1.0280221	0.2783134	-3.693757	0.2734862	-3.758954
A2	1.2112256	0.3876969	3.124156	0.3286096	3.685910
B2	-1.5960794	0.4548991	-3.508644	0.3533256	-4.517305
A3	0.8077615	0.2581445	3.129107	0.2732449	2.956181
B3	-0.9304350	0.2258458	-4.119781	0.2348299	-3.962165
GEE ,R[i]	**BETA**	**Cov1**	**Z Value**	**Cov2**	**Z value**
A1	0.38506904	0.1913023	2.0128828	0.1941331	1.9835308
B1	-0.10304272	0.2116799	-0.4867857	0.1949789	-0.5284814
A2	0.68818965	0.1526756	4.5075283	0.1870729	3.6787237
B2	0.12561936	0.2598234	0.4834798	0.1032378	1.2167967
A3	0.64637990	0.1499240	4.3113836	0.1534919	4.2111657
B3	-0.07108742	0.1499240	-0.3515459	0.1411093	-0.5037754

Tabel 5.Output program estimasi parameter data Trivariate pada N=500

GLM	**BETA**	**SE**	**Z Value**		
A1	0.9950	0.1124	8.855		
B1	-1.0129	0.1232	-8.219		
A2	0.9682	0.1245	7.775		
B2	-0.9420	0.1118	-8.424		
A3	0.8624	0.1205	7.159		
B3	-0.8723	0.1103	-7.911		
IEE ,R[i]=0	**BETA**	**Cov1**	**Z Value**	**Cov2**	**Z value**
A1	0.9949972	0.1124602	8.847546	0.1123722	8.854477
B1	-1.0129280	0.1237676	-8.184112	0.1232360	-8.219415
A2	0.9681825	0.1226742	7.892308	0.1245258	7.774958
B2	-0.9419733	0.1090755	7.892308	0.1118181	-8.424155
A3	0.8623992	0.1146893	-8.635973	0.1204695	7.158650
B3	-0.8723300	0.1041139	-8.378610	0.1102725	-7.910672
GEE ,R[i]	**BETA**	**Cov1**	**Z Value**	**Cov2**	**Z value**
A1	0.5711851	0.08470653	6.743106	0.09266715	6.163836
B1	-0.5431694	0.08198615	-6.625136	0.09512146	-5.710273
A2	0.2833776	0.08500234	3.333762	0.08525995	3.323689
B2	-0.5834261	0.08701083	-6.705213	0.09594521	-6.080826
A3	0.2386954	0.08147122	2.929813	0.07725042	3.089892
B3	-0.5729815	0.08403148	-6.818653	0.09361204	-6.120810

Tabel 6. Output program estimasi parameter data Trivariate pada N=1000

GLM	**BETA**	**SE**	**Z Value**		
A1	1.05861	0.08210	12.89		
B1	-1.09179	0.09255	-11.80		
A2	1.03207	0.09155	11.27		
B2	-1.00585	0.08353	-12.04		
A3	1.15401	0.09531	12.11		
B3	-1.09553	0.08722	-12.56		
IEE ,R[i]=0	**BETA**	**Cov1**	**Z Value**	**Cov2**	**Z value**
A1	1.058613	0.08463225	12.50838	0.08210276	12.89375
B1	-1.091786	0.09650697	-11.31303	0.09255138	-11.79654
A2	1.032070	0.09234371	11.17640	0.09154902	11.27341
B2	-1.005852	0.08389694	-11.98913	0.08353484	-12.04110
A3	1.154007	0.10054395	11.47764	0.09531109	12.10780
B3	-1.095532	0.09163161	-11.95583	0.08722218	-12.56024
GEE ,R[i]	**BETA**	**Cov1**	**Z Value**	**Cov2**	**Z value**
A1	0.43954202	0.09522404	4.615872	0.05830843	7.538224
B1	0.63575198	0.30641521	2.074806	0.04034376	15.758371
A2	-0.33368429	0.12093296	-2.759250	0.05947604	-5.610399
B2	0.68047896	0.29497742	2.306885	0.03728017	18.253109
A3	-0.09628762	0.09100109	-1.058093	0.05639416	-1.707404
B3	0.71589566	0.30606930	2.338999	0.03309670	21.630425

Tabel 7. Output program estimasi parameter data Trivariate pada N=2000

GLM	**BETA**	**SE**	**Z Value**		
A1	1.03655	0.05616	18.46		
B1	-0.93455	0.06022	-15.52		
A2	0.99456	0.06270	15.86		
B2	0.98483	0.05774	-17.06		
A3	1.00886	0.06374	15.83		
B3	-1.03518	0.05825	-17.77		
IEE ,R[i]=0	**BETA**	**Cov1**	**Z Value**	**Cov2**	**Z value**
A1	1.0365456	0.05607623	18.48458	0.05616112	18.45664
B1	-0.9345475	0.05904653	-15.82731	0.06021551	-15.52005
A2	0.9945644	0.06247027	15.92060	0.06269924	15.86246
B2	-0.9848259	0.05736188	-17.16865	0.05774204	-17.05561
A3	1.0088603	0.06423485	15.70581	0.06374447	15.82663
B3	-1.0351780	0.05913549	-17.50519	0.05824557	-17.77265
GEE ,R[i]	**BETA**	**Cov1**	**Z Value**	**Cov2**	**Z value**
A1	0.65181671	0.06394759	10.1929834	0.05045142	12.9196910
B1	-0.25681534	0.04850984	-5.2940872	0.04552572	-5.6411048
A2	0.02797271	0.08300896	0.3369842	0.04295013	0.6512835
B2	-0.38267349	0.05536828	-6.9114212	0.04383766	-8.7293329
A3	0.19297657	0.06320632	3.0531215	0.04383766	4.4428302
B3	-0.34316692	0.05744383	-5.9739561	0.04119977	-8.3293415