M - 7

RANDOM EFFECT MODEL AND GENERALIZED ESTIMATING EQUATIONS FOR BINARY PANEL RESPONSE

Jaka Nugraha
Dept. of Statistics, Islamic University of Indonesia, Kampus Terpadu UII, Jl. Kaliurang Km.14, Yogyakarta, Indonesia

email: jk.nugraha@gmail.com

Abstract

Panel data models are widely used in empirical analysis because they allow researchers to control for unobserved individual time-invariant characteristics. However, these models pose important technical challenges. In particular, if individual heterogeneity is left completely unrestricted, and then estimates of model parameters in nonlinear and/or dynamic models suffer from the incidental parameters problem. This problem arises because the unobserved individual characteristics are replaced by inconsistent sample estimates, which, in turn, biases estimates of model parameters. Logit model or probit model on panel data with using univariate approximation (neglect correlation) result consistent estimator but not efficient. In many cases, data are multivariate or correlated (e.g., due to repeated observations on a study subject or for subjects within centers) and it is appealing to have a model that maintains a marginal logistic regression interpretation for the individual outcomes.
In this paper, we studied modeling binary panel response using Random Effects Model (REM). Using Monte Carlo Simulation, we research correlations effects to maximum likelihood estimator (MLE) of random effects model. We also compare MLE of REM to Generalized Estimating Equations (GEE) of logit model. Data were generated by using software R.2.8.1 as well as the estimation on the parameters. Based on the result, it can be concluded that (a) In some value of individual effect, random effects model is more better GEE. (b) REM can be accommodating individual effects and closer to parameter than the other. (c) REM is appropriate method to estimate covarians of utility at individual effect having value about one.

Keywords : Random Utility Models, Maximum Likelihood Estimator Generalized Estimating Equations, Logit Models, Probit Models

1. INTRODUCTION

Panel data models are widely used in empirical analysis because they allow researchers to control for unobserved individual time-invariant characteristics. However, these models pose important technical challenges. In particular, if individual heterogeneity is left completely unrestricted then estimates of model parameters in nonlinear and/or dynamic models suffer from the incidental parameters problem. This problem arises because the unobserved individual characteristics are replaced by inconsistent sample estimates, which, in turn, biases estimates of model parameters (Greene, 2003). Liang and Zeger (1986), shown that Logit model or Probit

model on panel data with using univariate approximation (neglect correlation) result consistent estimator but not efficient. In many cases, data are multivariate or correlated (e.g., due to repeated observations on a study subject or for subjects within centers) and it is appealing to have a model that maintains a marginal logistic regression interpretation for the individual outcomes. Commonly used logistic random effects models do not have this property, since the logistic structure is lost in integrating out the random effects. An alternative is to use a marginal analysis that avoids complete specification of the likelihood (Liang and Zeger, 1986; Prentice, 1988; Lipsitz *et al.*). Prentice (1988) proposed modeling strategic by GEE to obtain consistent estimator and normal asymptotic. GEE are hindered multiple integral by marginal distribution.

Nugraha *et al.*(2008) have tested Logit Model in multivariate binary response using Monte Carlo simulation. They concluded that GEE more proper on height correlation, although estimators of correlation was underestimated. MLE of Probit Model could not derive by analytic, because the likelihood functions formed a multiple integral. Simulation approximation to compute multiple integral caused bias. The others problem on Probit Model are the log likelihood function not global concave, so there are no one solutions. Simulation methods rely on approximating an integral (that does not have a closed form) through Monte Carlo integration. Draws are taken from the underlying distribution of the random variable of integration and used to calculate the numeric integral. Simulated maximum log likelihood estimation is a common estimator used for Probit Model and random effects model. Such estimators exhibit a non-negligible bias when too few draws are used in estimation, and prior research exists regarding the magnitude and properties of this bias with respect to quasi-random draws (Bhat, 2001). Train (1999) provide further evidence of the benefits of intelegent drawing techniques such as Halton and Shuffled Halton, which require fewer numbers of draws than pseudo-random in order to uncover identification issues.

In this paper, we studied modeling binary panel response using random effects model. Using Monte Carlo Simulation, we research correlations effects to maximum likelihood estimator (MLE) of Random Effects model. We also compare MLE on Random Effects model to GEE on Logit Model.

2. GEE MODEL

In the panel response within exponential family distribution, Liang dan Zeger (1986) proposed the GEE model. For the binary response ($Y_{i1},...,Y_{iT}$) with each Y_{it} binary value (dichotomous), both link logit and link probit can be utilized for GEE model. Contoyannis et al. (2001) has been constructed probit model on binary panel data by the model of :

$$Y_{it} = \beta' X_{it} + \xi_i + \varepsilon_{it} \quad \text{for } i=1,...,n \text{ and } t=1,...,T \qquad (1)$$

ξ_i is individual effect within the normal distribution and mean value of null and the variance of σ_η^2. Whereas $\varepsilon_{it} \sim N(0, \sigma^2)$ and independent with ξ_i. X_{it} are vector px1 respect to the independent variable for responden i on t periode. β is a parameter vector in px1 size. Individual probability i for making a decision series was calculated by using conditional probability.

We assume that each of n individual observed T times. Y_{it} is t^{nd} response on i^{nd} individual/subject and each response are binary. So, response for i^{nd} individual can be

$$Y_i = (Y_{i1},....,Y_{iT})$$

that is vector 1xT. $Y_{it} = 1$ if i^{nd} subject and t^{nd} response choose the first alternative and $Y_{it} = 0$ if choose the second alternative. Each subject have covariate X_i (individual characteristic) dan covariate Z_{ijt} (characteristic of alternative j=0,1). To simplify, we choose one of individual characteristic and one of characteristic of alternative. Utility of subject i choose alternative j on

response t is

$$U_{ijt} = V_{ijt} + \varepsilon_{ijt} \text{ . for } t=1,2,...,T \text{ ; } i=1,2,...,n \text{ ; } j=0,1 \quad (2)$$

with $V_{ijt} = \beta 0_{jt} + \beta_{jt}X_i + \gamma_t Z_{ijt}$.

By assumption that decision makers choosing alternative based on maximum utility, model can be represented in different of utility,

$$U_{it} = V_{it} + \varepsilon_{it} \quad (3)$$

with $V_{it} = (V_{i1t} - V_{i0t})$ and $\varepsilon_{it} = (\varepsilon_{i1t} - \varepsilon_{i0t})$.

Expectation of eq. (3) are

$E(\varepsilon_{it}) = E(\varepsilon_{i1t}) - E(\varepsilon_{i0t}) = 0.5772 - 0.5772 = 0$

$E(\alpha_i) = 0$; $E(U_{it}) = V_{it}$

and theirs varians are

$$Var(\varepsilon_{it}) = Var(\varepsilon_{i1t}) + Var(\varepsilon_{i0t}) = \frac{\pi^2}{6} + \frac{\pi^2}{6} = \frac{\pi^2}{3}$$

$$Var(\alpha_i) = Var(\alpha_{i0}) + Var(\alpha_{i1}) = 2\sigma^2$$

$$Var(U_{it}) = Var(\alpha_i) + Var(\varepsilon_{it}) = 2\sigma^2 + \frac{\pi^2}{3}$$

Covariance and Correlation among utilities are

$Cov(U_{it};U_{is}) = Cov((\alpha_i + \varepsilon_{it}), (\alpha_i + \varepsilon_{is})) = 2\sigma^2$ for all $t \neq s$

$$Cor(U_{it};U_{is}) = \frac{2\sigma^2}{\left(2\sigma^2 + \frac{\pi^2}{3}\right)}$$

Probability of subject i choose ($y_{i1} = 1,...., y_{iT} = 1$) is

$$P(y_{i1} = 1,...., y_{iT} = 1) = \int_{\varepsilon i} I(-V_{it} < \varepsilon_{it}).f(\varepsilon_i)d\varepsilon_i \qquad \forall t \quad (4)$$

This probability value is multiple integral and depending on parameters β, γ and distribution of ε (Train, 2003).

The logit model can be derived by assumption that ε_{ijt} have Extreme Value Type I distribution (Gumbel) and independence each other (all i, j and t). Probability of subject i choose j=1 for response t^{nd} is

$$P(y_{it} = 1) = \pi_{it} = \frac{\exp(V_{i1t})}{[\exp(V_{i0t}) + \exp(V_{i1t})]} = \frac{\exp(V_{it})}{[1 + \exp(V_{it})]} \quad (5)$$

with

$$V_{ijt} = \beta 0_{jt} + \beta_{jt}X_i + \gamma_t Z_{ijt} \quad \text{for } t=1,2,...,T \text{ ; } i=1,2,...,n \text{ ; } j=0,1.$$

On Logit Model, GEE are easier to implement than MLE. GEE use approximation by marginal distribution and can be represented by

$$G(\theta) = \sum_{i=1}^{n} W_i \Delta_i S_i^{-1}(Y'_i - \pi'_i) = 0 \quad (6)$$

$$\text{with } W_i = diag\begin{pmatrix} 1 & 1 \\ X_i & .... & X_i \\ (Z_{i11} - Z_{i01}) & (Z_{i1T} - Z_{i0T}) \end{pmatrix} ; \Delta_i = diag\left(\pi_{i1}(1-\pi_{i1}) \quad ... \quad \pi_{iT}(1-\pi_{iT})\right)$$

$$S_i = A_i^{1/2} R_i A_i^{1/2} \text{ with } A_i^{1/2} = \begin{pmatrix} \sqrt{Var(Y_{i1})} & \ldots & 0 \\ \ldots & \ldots & \ldots \\ 0 & \ldots. & \sqrt{Var(Y_{iT})} \end{pmatrix}$$

with $Y_i = (Y_{i1},....,Y_{iT})$; $\pi_i = (\pi_{i1},...., \pi_{iT})$. Estimators GEE are solving equations (6) on sample data W (Nugraha et al., 2008) .

GEE on Probit Model are solving of estimating equation

$$G(\theta) = \sum_{i=1}^{n} W_i \Delta_i S_i^{-1} (Y'_i - \pi'_i) = 0 \qquad (7)$$

with $\pi_{it} = \Phi(V_{it})$; $\Delta_i = diag(\phi(V_{it}))$. Estimations of parameter correlations are underestimated. GEE on probit model are equivalent to GEE on Logit Model.

3. RANDOM EFFECTS MODEL

From the equation of utility difference (3), we added the individual effect α_i

$$U_{it} = V_{it} + \alpha_i + \varepsilon_{it} \qquad (8)$$

α_i is effect of individual i having normal distribution, $\alpha_i \sim NID(0, \sigma^2)$ and independent to ε_{it}. ε_{it} have Extreme Value Distribution.

Based on equation (8), we will estimate parameters (σ, β_t, γ_t) for t=1,..,T. By equation (5), we have conditional probability :

$$g_{it} = P(y_{it}=1|\alpha_i) = P(\varepsilon_{it} < -V_{it}|\alpha_i) = \pi_{it} \mid \alpha_i = \frac{\exp(V_{it}+\alpha_i)}{[1+\exp(V_{it}+\alpha_i)]}$$

$$= \frac{\exp(V_{it}+\alpha_i)}{[1+\exp(V_{it}+\alpha_i)]} \qquad (9)$$

Marginal probabilities from equation (4) for Random Effect Model are

$$P(y_{it}=1) = \pi_{it} = \int_{-\infty}^{\infty} P(y_{it} \mid \alpha_i) f(\alpha_i) d\alpha_i$$

$$= \int_{-\infty}^{\infty} \frac{\exp(V_{it}+\alpha_i)}{[1+\exp(V_{it}+\alpha_i)]} \phi(\alpha_i) d\alpha_i \qquad (10)$$

$\phi(\xi_{it})$ is standard normal density.

$$P(y_{i1}=1, \ldots, y_{iT}=1) = \int_{-\infty}^{\infty} \prod_{t=1}^{T} P(y_{it} \mid \alpha_i) f(\alpha_i) d\alpha_i$$

$$= \int_{-\infty}^{\infty} \prod_{t=1}^{T} \frac{\exp(V_{it}+\alpha_i)}{[1+\exp(V_{it}+\alpha_i)]} \phi(\alpha_i) d\alpha_i \qquad (11)$$

So,

$$P(y_{i1}, \ldots, y_{iT}) = \int_{-\infty}^{\infty} \prod_{t=1}^{T} \left(\frac{\exp(V_{it}+\alpha_i)}{[1+\exp(V_{it}+\alpha_i)]} \right)^{y_{it}} \left(1 - \frac{\exp(V_{it}+\alpha_i)}{[1+\exp(V_{it}+\alpha_i)]} \right)^{1-y_{it}} \phi(\alpha_i) d\alpha_i$$

MLE of parameters (β_t, γ_t ;σ), t=1,..T, can be obtained from likelihood function :

$$L(\beta,\gamma,\sigma)=\prod_{i=1}^{n}\int_{-\infty}^{\infty}\left(\prod_{t=1}^{T}(g_{it})^{y_{it}}(1-g_{it})^{1-y_{it}}\right)\phi(\alpha_i)\partial\alpha_i \quad (12)$$

Then the log-likelihood function is

$$LL=\log L(\beta,\gamma,\sigma)=\sum_{i=1}^{n}\log\left(\int_{-\infty}^{\infty}\prod_{t=1}^{T}\left((g_{it})^{y_{it}}(1-g_{it})^{1-y_{it}}\phi(\alpha_i)\right)\partial\alpha_i\right) \quad (13)$$

There are some methods of iteration to get the solution of equatian (13) : Newton-Raphson methods, BFGS methods, etc.

4. GENERATING DATA SIMULATION AND RESULT

We will generate simulation data with T=3, n=1000. Then, the equations of utility are

$U_{i0t} = \alpha_{i0} + \beta 0_{0t} + \beta_{0t}X_i + \gamma_t Z_{i0t} + \varepsilon_{i01}$ and

$U_{i1t} = \alpha_{i1} + \beta 0_{1t} + \beta_{1t}X_i + \gamma_t Z_{i1t} + \varepsilon_{i11}$ (14)

for i=1,...,N; j=0,1 and t=1,...,3; ε_{ijt} ~Extreme Value Type I, $\alpha_i \sim N(0,\sigma^2)$.

Equation (14) can be presented in difference of utility $U_{it} = U_{i1t} - U_{i0t}$. On Logit Model, equations of utility difference are

$U_{it} = \beta 0_t + \beta_t X_i + \gamma_t Z_{it} + \alpha_i + \varepsilon_{it}$ (15)

or

$U_{i1} = V_{i1} + \alpha_i + \varepsilon_{i1}$; $U_{i2} = V_{i2} + \alpha_i + \varepsilon_{i2}$; $U_{i3} = V_{i3} + \alpha_i + \varepsilon_{i3}$

with

$V_{it} = (V_{i1t} - V_{i0t}) = \beta 0_t + \beta_t X_i + \gamma_t Z_{it}$.

$Z_{it} = (Z_{i1t} - Z_{i0t})$; $\beta 0_t = \beta 0_{0t} - \beta 0_{1t}$; $\beta_t = \beta_{0t} - \beta_{1t}$; $\alpha_i = (\alpha_{i1} - \alpha_{i0})$.

We generate data on $\beta 0_t = -1$; $\beta_t = 0.5$, $\gamma_t = 0.3$ and some of variance σ^2 =0; 0.5; 1; 2; 4, 6 using program R.2.8.1. From the data simulation, we built the Logit Model using three approximations. First, we assume independence each other for all i, j and t (independent logit model) and estimate parameters using MLE (θ_{MLECS}) . Second, we estimate parameter using GEE (θ_{GEE}). The last approximation is Random Effect Model using MLE (θ_{MLERE}). Results of the simulations presented in Figure 1 to Figure 10.

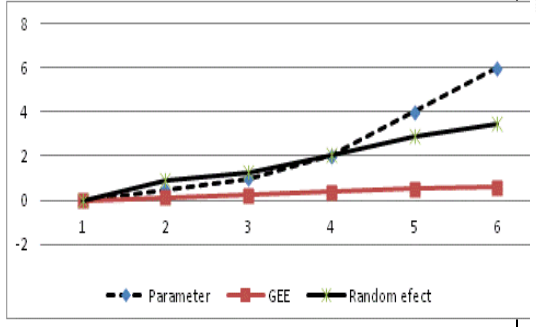

Figure 1. Estimator of σ^2

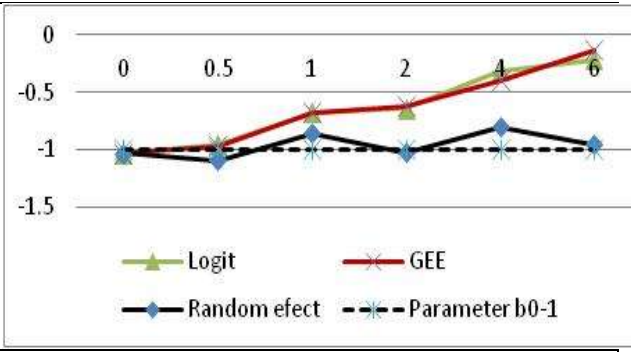

Figure 2. Estimator of $\beta 0_1$

Based on simulation result representing at Figure 1., estimator of σ^2 using Random effect Model (θ_{MLERE}) more accurate than GEE (θ_{GEE}). θ_{MLERE} have value closer to real value of parameter than θ_{GEE}.

Overall, estimator of $\beta 0_t$, β_t and γ_t in MLE as same as GEE (see Figure 2 to Figure 10).

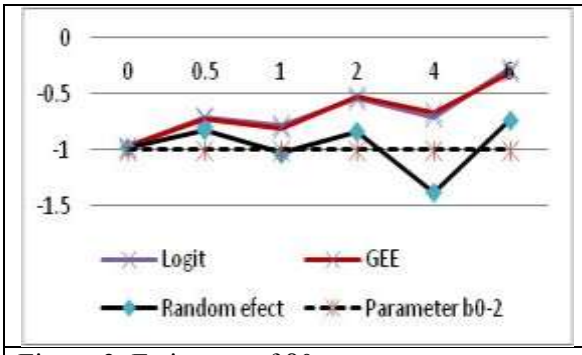

Figure 3. Estimator of $\beta 0_2$

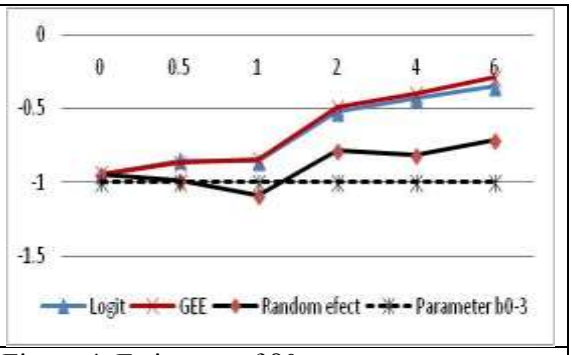

Figure 4. Estimator of $\beta 0_3$

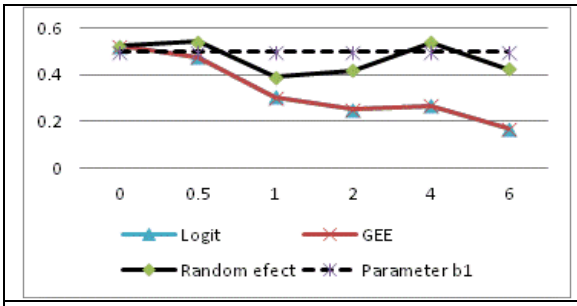

Figure 5. Estimator of β_1

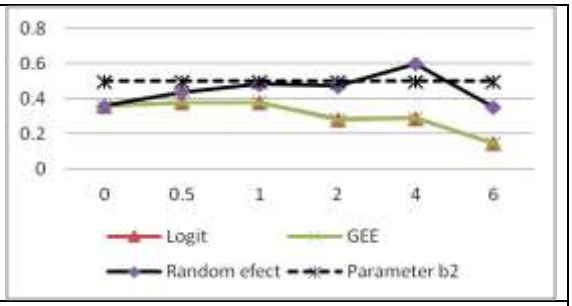

Figure 6. Estimator of β_2

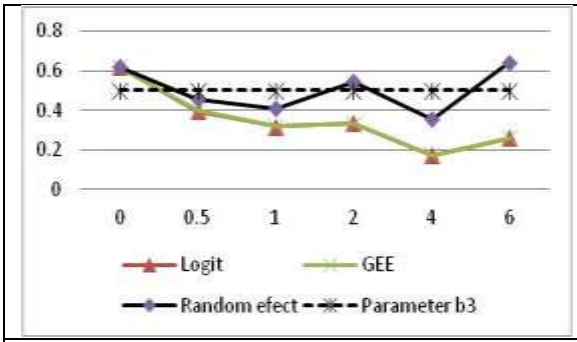

Figure 7. Estimator of β_3

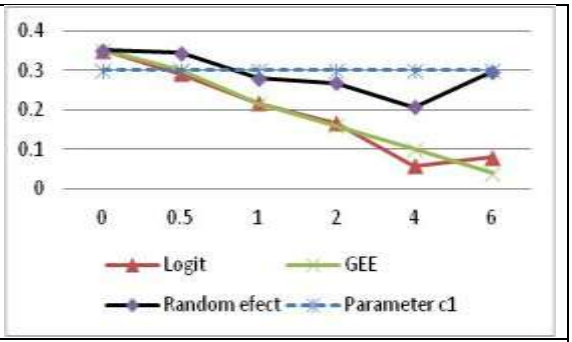

Figure 8. Estimator of γ_1

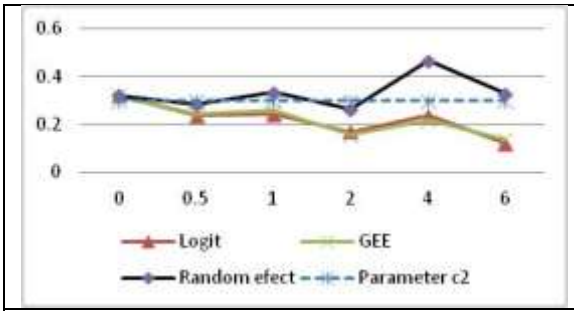

Figure 9. Estimator of γ_2

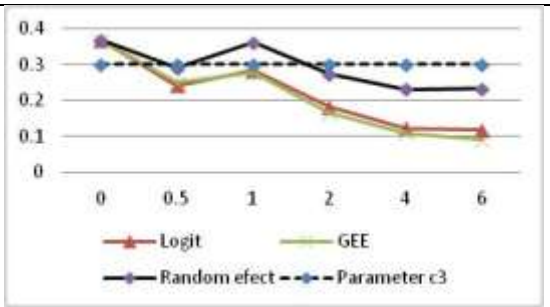

Figure 10. Estimator of γ_3

We can see that on $\sigma^2 = 0$ (no effect individual) estimators by three approximations are same.

On general, estimator of independent Logit Model ($\hat{\theta}_{MLECS}$) and estimator of GEE ($\hat{\theta}_{GEE}$ **)** are not different but increasing of σ^2 impact to increasing bias of estimator. On data having individual effect (see Figure 2 to Figure 10), the random effect model ($\hat{\theta}_{MLERE}$ **)** better than $\hat{\theta}_{MLECS}$ and $\hat{\theta}_{GEE}$ **.**

By GEE, we estimate coefficient regressions and correlation among alternative. Using effect random model we can estimate coefficient regressions, individual effect and covarians/correlations among alternative. On value of individual effect σ^2 about 1, random effect model can estimate covarians of utility more appropriate than the others value of σ^2.

5. CONCLUSION

On modelling binary panel response, we can use Random Effects Model. This model use the Extreme Value distribution and the standart normal distribution. The model is

$$P(y_{i1}, \ldots, y_{iT}) = \int_{-\infty}^{\infty} \prod_{t=1}^{T} (g_{it})^{y_{it}} (1-g_{it})^{1-y_{it}} \phi(\alpha_i) d\alpha_i$$

with $g_{it} = \dfrac{\exp(V_{it}+\alpha_i)}{[1+\exp(V_{it}+\alpha_i)]}$. The log-likelihood function is

$$\log L(\beta,\gamma,\sigma) = \sum_{i=1}^{n} \log\left(\int_{-\infty}^{\infty} \prod_{t=1}^{T} \left((g_{it})^{y_{it}} (1-g_{it})^{1-y_{it}} \phi(\alpha_i) \right) \partial\alpha_i \right)$$

Based on simulation, Random Effects Model more appropriate than GEE.

REFERENCE

Bhat, C. (2001), 'Quasi-random maximum simulated likelihood estimation of the mixedmultinomial logit model', *Transportation Research B* **35**, 677–693

Contoyannis P, Andrew M. J, Rice N (2001), Dynamics of Health in British Household: Simulation-Based Inference in Panel Probit Model, *Working Paper*, Department of Economics and Related Studies, University of York

Greene W. (2003), ***Econometrics Analysis***, 5 Editions, Prentice Hall

Liang, K.Y., dan Zeger,S.L (1986). 'Longitudinal Data Analysis Using Generalised Linear Models', *Biometrika 73, 13-22.*

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*

Nugraha J. (2008), 'Estimatiom Probit Model on Multivariate Biner Response Using MLE and GEE', National Seminar at Dept of Math UGM 31 Mei 2008

Nugraha J, Haryatmi S. and Guritno S, (2008), 'Modeling Distribution Extreme Value On Multivariate Binary Response Using MLE and Quasi MLE', 7th World Congress in

Probability and Statistics at the National University of Singapore from 14 to 19 July, 2008

Prentice (1988), 'Correlated Binary Regression with Covariates Specific to Each Binary Observation'. *Biometrics 44, 1043-1048.*

Train, K. (1999), 'Halton Sequences for Mixed Logit. University of California, Berkeley.

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