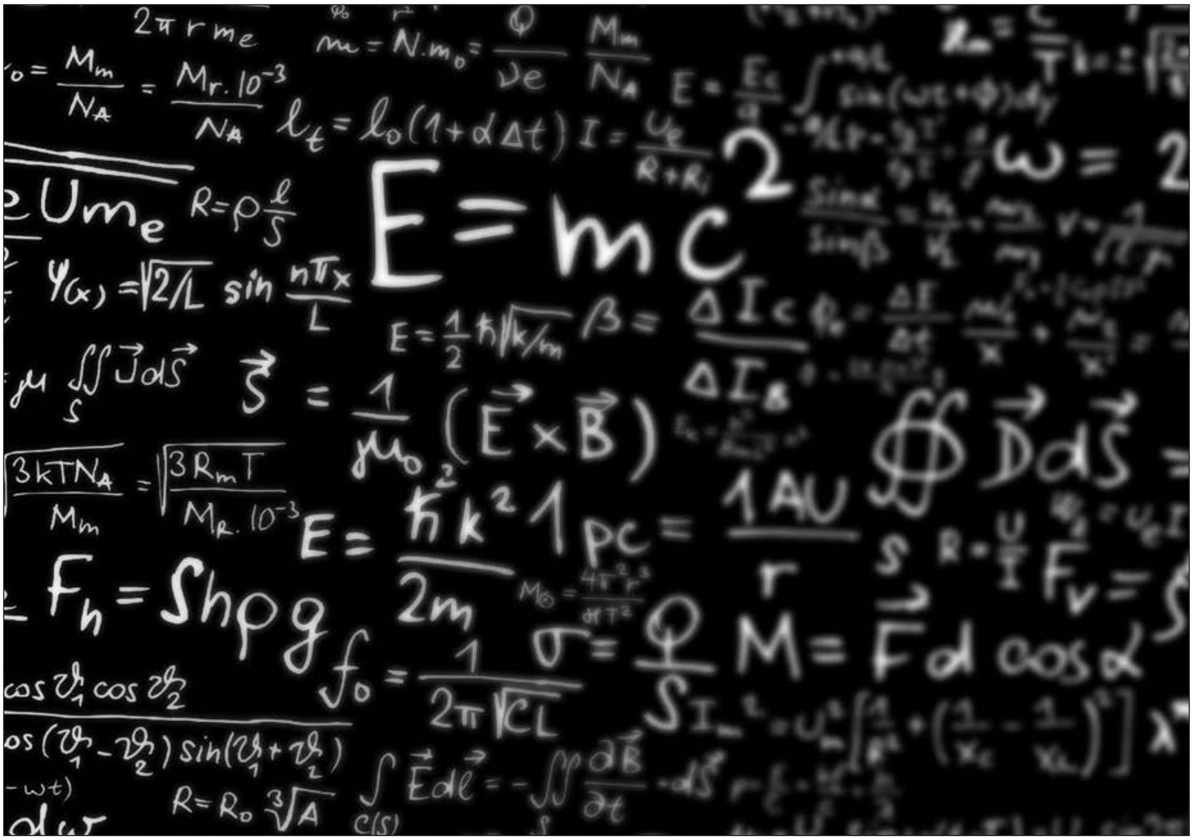

Asian Journal of Mathematics & Statistics

ISSN 1994-5418

science alert

ANSInet
an open access publisher
http://ansinet.com

Asian Journal of Mathematics and Statistics 4 (3): 113-127, 2011
ISSN 1994-5418 / DOI: 10.3923/ajms.2011.113.127

Mixed Logit Model on Multivariate Binary Response using Maximum Likelihood Estimator and Generalized Estimating Equations

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

ABSTRACT

This study presents discussion on the effects of correlation among response respect to estimator properties in mixed logit model on multivariate binary response. It is assumed that each respondent was observed for T response. Y_{it} is the t^{th} response for the i^{th} individual/subject and each response is binary. Each subject has covariate X_i (individual characteristic) and covariate Z_{ijt} (characteristic of alternative j). Individual response i that is represented by $Y_i = (Y_{i1},....,Y_{iT})$, Y_{it} is t^{nd} response on i^{th} individual/subject and the response is binary. In order to simplify, one of individual characteristic was and alternative characteristics. We studied effects of correlations using data simulation. Methods of estimations used in this study are Generalized Estimating Equations (GEE) and Maximum Likelihood Estimator (MLE). We generate data and estimate parameters using software R.2.10. From simulation data, we conclude that MLE on mixed logit model is better than GEE. The higher correlation among utility, the higher deviation estimator to parameter.

Key words: Random utility model, simulated maximum likelihood estimator, GEE, GHK simulator, Newton-Raphson

INTRODUCTION

Discussion on binary response modeling was reported by Sarkar and Midi (2010), Zadkarami (2008), Sarkar *et al.* (2011) and the models usually adopted probit model and logit model (Alpu and Fidan, 2004; Damisa *et al.*, 2007). Generally for panel data observations towards same subjects and variables are conducted repeatedly. The model was constructed based on approach of fixed effect, random and dynamic models (Lechner *et al.*, 2005). Simulation study of an efficient estimation of a mean true effect using a fixed effects model and a random effects model in order to establish appropriate confidence interval estimation have been reported (Antoine *et al.*, 2007). Method for estimating parameter on panel data modeling used was Maximum Likelihood Estimating (MLE), moment method and Generalized Estimating Equation (GEE) (Diggle *et al.*, 1996). The simplest model is independent model which the assumption taken based on the response on the same and inter subjects is independent. Based on this assumption, combined probability is result of multiplication of its marginal probability. Liang and Zeger (1986) suggest that logistic and probit analysis that are ignoring the correlation and produce consistence estimation parameter but still have close relation becomes inefficient. Chaganty and Joe (2004) have proposed the use of equations of the correlation estimator for probit models in GEE.

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. An alternative is to use a marginal analysis that avoids complete specification of the likelihood (Lipsitz *et al*., 1990). Prentice (1988) demonstrate modeling strategic to obtain a consistent and normal asymptotic regression coefficient estimator by using GEE. Double regression was not needed in this calculation.

Development of model by controlling unobserved and homogeneous individual characteristic towards replication was reported. However, for heterogeneous individual characteristic, problem would be found is that estimation parameter becomes bias (Greene, 2005). Double integral was needed in probit model. From several simulations, Hajivassiliou *et al.* (1996) reported that Geweke-Hajivasilou-Keane (GHK) was the best simulation. Geweke *et al.* (1997) also find double integral calculation by using Monte Carlo simulation known as unbias and consistence GHK.

Eventually, several different independent variables are observed simultaneously. This observation produces multivariate data. Nugraha (2000) examined properties of parameter estimators on bivariate logistic regression by using MLE and GEE method. These method produce a consistence estimator. Nugraha *et al.* (2006) reported that for logistic model on multivariate binary, smaller variance estimator gained by GEE and compared to independent assumption approach. Nugraha *et al.* (2009a) evaluated the comparison of GEE and MLE method for multivariate binary logistic. It was concluded that GEE model has smaller bias. However, the estimator for correlation parameter is always underestimate. Akanda *et al.* (2005) proposed an alternative procedure for goodness of fit test based on GEE where, the correlation between two responses was considered.

Nugraha *et al.* (2008) presented that analytically, both first and second derivation of log-likelihood binary multivariate model can be calculated. Based on its first derivation and second derivation, estimation on the parameter can be executed by Newton-Raphson iteration procedure. Nugraha *et al.* (2009b) obtained the estimator of binary multivariate probit model by Simulated Maximum Likelihood Estimator (SMLE) based on GHK simulation.

During the last decade, practitioners and researchers are willing to move from the multinomial and nested logit models, that were the standard until only a few years ago, towards these more general models, specifically towards Mixed Logit (Munizaga and Alvarez-Daziano, 2004). In this present study, correlation effect on mixed logit model towards binary multivariate data is discussed. Estimation of parameter is conducted by MLE and GEE method and comparative study for simulation data is subjected. Both simulation data and estimation of parameter were performed utilizing R.2.10.0. program. (www.R-project.org).

MULTIVARIATE BINARY RESPONSE

It is assumed that Y_{it} is binary response, $Y_{it} = 1$ as the subject i at the response of t choosing the alternative 1 and $Y_{it} = 0$ if the subject of i at the response of t choosing the alternative of 2. Each individual has covariate X_i as individual characteristics i and covariate Z_{ijt} as characteristic of choice/alternative j at the individu of i.

Utility of subject i selecting the alternative of j on response t is:

$$U_{ijt} = V_{ijt} + \varepsilon_{ijt} \text{ for } t = 1, 2, \ldots, T;\ i = 1, 2, \ldots, n;\ j = 0, 1 \tag{1}$$

with:

$$V_{ijt} = \alpha_{jt} + \beta_{jt} X_i + \gamma_t Z_{ijt}$$

By assuming that decision maker select the alternative based on the maximum utility value, the model can be expressed in the form of difference utility:

$$U_{it} = U_{i1t}\text{-}U_{i0t} = V_{it}+\varepsilon_{it} \tag{2}$$

with:

$$V_{it} = (V_{i1t}-V_{i0t}) = (\alpha_{1t}\text{-}\alpha_{0t})+(\beta_{1t}\text{-}\beta_{0t})\,X_i+\gamma_t\,(Z_{i1t}\text{-}Z_{i0t}) = \alpha_t+\beta_t X_i+\gamma_t Z_{it}$$

and $\varepsilon_{it} = (\varepsilon_{it}\text{-}\varepsilon_{i0t})$. The probability of subject of i selecting ($y_{i1} = 1, ..., y_{iT} = 1$) is:

$$P\,(y_{i1} = 1, ..., y_{iT} = 1) = \int_{\varepsilon i} I\,(-V_{it}<\varepsilon_{it}).\ f\,(\varepsilon_i)\ d\varepsilon_{i1}...\ d\varepsilon_{iT}\ \forall_t \tag{3}$$

With $\varepsilon_i = (\varepsilon_{i1}, ..., \varepsilon_{iT})'$. Value of probability is calculated by multiple integral T depend on the parameter $\theta = (\theta_1', \theta_2', ..., \theta_T')$ as well as distribution of ε (Train, 2003).

GEE ON MODEL LOGIT

GEE on multivariate binary response can be expressed in the following form :

$$G(\theta) = \sum_{i=1}^{n} W_i \Delta_i S_i^{-1}(Y'_i - \pi'_i) = 0 \tag{4}$$

W_i is called as observation matrix:

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

and $\Delta_i = \text{diag}\,(\pi_{i1}\,(1\text{-}\pi_{i1})\ ...\ \pi_{iT}\,(1\text{-}\pi_{iT}))$ S_i is matrix T×T and can be expressed as:

$$S_i = A_i^{½}\,R_i\,A_i^{½}$$

It is defined that $A_i^{1/2} = \text{diag}\left(\sqrt{\text{Var}(Y_{i1})}\ \ \cdots\ \ \sqrt{\text{Var}(Y_{iT})}\right)$ and R_i is "working" matrix with the correlation Y_i in the size of T×T.

Paired correlation coefficient between response of s and response of r, ρ_{isr}= corr (Y_{is}, Y_{ir}) for i = 1, 2, ..n and s, r = 1, 2, ... T. In order to estimate R_i, empirical correlation, r_i is defined as vector with the size of T (T-1)/2 with the elements:

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

Empirical correlation r_{ist} that is contained parameter π_{is} and π_{it} is unbias estimator for ρ_{ist} (Liang and Zeger, 1986). As it is assumed that $\rho_{ist} = \rho_{st}$ for all of i, the estimator is:

$$\hat{\rho}_{st} = \frac{1}{n}\sum_{i=1}^{n} r_{ist} \tag{5}$$

Equation 4 and 5 can be solved at the same time to get the parameter estimator θ and ρ.

Theorem 1: If $\hat{\theta}_G$ is GEE estimator that is solution of Eq. 4, so $\lim_{n\to\infty}\sqrt{n}(\hat{\theta}_G - \theta)$ is normal multivariate distributed:

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

where,

$$\mathrm{Var}\,(\hat{\theta}_G) = \lim_{n\to\infty} n\left(\sum_{i=1}^{n}\left(\frac{\partial\hat{\pi}_i}{\partial\theta}\right)S_i^{-1}\left(\frac{\partial\hat{\pi}_i}{\partial\theta}\right)'\right)^{-1}\left(\sum_{i=1}^{n}\left(\frac{\partial\hat{\pi}_i}{\partial\theta}\right)\Sigma_i^{-1}[\mathrm{Var}(y'_i)]S_i^{-1}\left(\frac{\partial\hat{\pi}_i}{\partial\theta}\right)'\right)\left(\sum_{i=1}^{n}\left(\frac{\partial\hat{\pi}_i}{\partial\theta}\right)S_i^{-1}\left(\frac{\partial\hat{\pi}_i}{\partial\theta}\right)'\right)^{-1}$$

MIXED LOGIT MODEL

Mixed logit model is combination of extreme value distribution and normal distribution assumption. Based on equation of utility difference (2), the individual effect δ_{it} is added:

$$U_{it} = V_{it}+\delta_{it}+\varepsilon_{it} \tag{6}$$

ε_{it} has extreme value type I distribution. δ_{it} is ith individual effect of tth response and is independent respect to ε_{it}. It is assumed that $\delta_i = (\delta_{i1}, \ldots, \delta_{iT})$ has multivariate normal, $\delta_{it} \sim N$ (0, 1) and $\delta_i \sim N$ (0, Σ). Parameter that will be estimated on this mixed Logit model are the parameter of regression coefficient (α_t, β_t, γ_t) and Σ. Parameter Σ are estimated by the Cholesky factor, C.

Individual effect δ_i can be can be modified in the following form:

$$\delta_i = C\xi_i;\ \xi_i = (\xi_{i1}, \ldots, \xi_{iT})' \tag{7}$$

where, C is matrix of Cholesky factor of Σ:

$$CC' = \Sigma \text{ and } \xi_i \sim N\ (0, 1)$$

I is identity matrix with the size of T×T. Generally, for t = 1, ... , T, the utility function can be constructed:

$$U_{it} = V_{it}+K_{it}+\varepsilon_{it} \tag{8}$$

where,

$$K_{it} = \sum_{l=1}^{t} c_{tl}\,\xi_{il},\ x_{it} \sim \mathrm{NID}\ (0, 1)$$

$\varepsilon_{it} \sim$ extreme value type I, c_{tl} is an element of cholesky matrix. Based on Eq. 8 conditional logit model is derived with expression:

$$g_{it} = P\ (y_{it} = 1|K_{it}) = \frac{\exp\ (V_{it} + \sum_{l=1}^{t} c_{tl}\xi_{il})}{[1 + \exp\ (V_{it} + \sum_{l=1}^{t} c_{tl}\xi_{il})]} \tag{9}$$

Marginal probability (for t and i) in the mixed logit model is:

$$\begin{aligned} P(y_{it} = 1) = \pi_{it} &= \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} P(y_{it} \mid c) f\ (\xi_{i1}, .., \xi_{it}) d\xi_{i1}\ \ldots\ d\xi_{it} \\ &= \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \frac{\exp\ (V_{it} + \sum_{l=1}^{t} c_{tl}\xi_{il})}{[1 + \exp\ (V_{it} + \sum_{l=1}^{t} c_{tl}\xi_{il})]} \prod_{l=1}^{t} \varphi(\xi_{il}) d\xi_{i1} \ldots d\xi_{it} \end{aligned} \tag{10}$$

ϕ (ξ_{it}) is standard normal density.

The joint conditional probability is:

$$P\left(Y_{i1} = y_{i1}, \ldots, Y_{iT} = y_{iT} \mid (c, \xi_i)\right) = \prod_{t=1}^{T} (g_{it})^{y_{it}} (1 - g_{it})^{1 - y_{it}} \tag{11}$$

where, c = (c_{11}, c_{12}, c_{13}, c_{22}, c_{23}, c_{33}) and ξ_i (ξ_{i1}, ..., ξ_{iT}). The joint probability is:

$$P\left(Y_{i1} = y_{i1}, \ldots, Y_{iT} = y_{iT})\right) = \int_{-\infty}^{\infty} .. \int_{-\infty}^{\infty} \left(\prod_{t=1}^{T} (g_{it})^{y_{it}} (1 - g_{it})^{1 - y_{it}} \right) \varphi(\xi_i) d\xi_{i1} \ldots d\xi_{iT} \tag{12}$$

ϕ (ξ_i) is normal standard density.

Theorem 2: If it is known the utility model as represented in Eq. 8 that is fit to regularity condition, so MLE for parameter of $\theta_t = (\alpha_t, \beta_t, \gamma_t)'$ and $\sum$ is the solution of estimator equation:

$$\sum_{i=1}^{n} \begin{pmatrix} 1 \\ X_i \\ Z_{it} \end{pmatrix} (y_{it} - \frac{Q_1}{Q}) = 0 \tag{a}$$

$$\frac{1}{Q} \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \prod_{s=1}^{T} (g_{is})^{y_{is}} (1 - g_{is})^{1 - y_{is}} (y_{it} - g_{it}) \xi_{il}\ \varphi(\xi_i) d\xi_{i1} \ldots d\xi_{iT} = 0 \tag{b}$$

where, l = 1, ..., T. with:

$$Q = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \prod_{s=1}^{T} \left((g_{is})^{y_{is}} (1 - g_{is})^{1 - y_{is}} \right) \varphi(\xi_i) d\xi_{i1} \ldots d\xi_{iT}$$

$$Q_1 = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \prod_{s=1}^{T} (g_{is})^{y_{is}} (1 - g_{is})^{1 - y_{is}}\ g_{it} \varphi(\boldsymbol{\xi}_i) d\xi_{i1} \ldots d\xi_{iT}$$

$\hat{\theta}_{MLE}$ and $\hat{\Sigma}_{MLE}$ can be obtained by using iteration process of methods BHHH ((Berndt, Hall, Hall, Hausman) or BFGS (Broyden, Fletcher, Goldfarb dan Shanno) (Chong and Zak, 1996).

SIMULATION DATA

Simulation data was simulated by taking case T = 3 and n = 1000. Utility equation is:

$$U_{i0t} = \alpha_{0t}+\beta_{0t} X_i+\gamma_t Z_{i0t}+\varepsilon_{i01} \text{ dan } U_{i1t} = \alpha_{1t}+\beta_{1t} X_i+\gamma_t Z_{i1t}+\varepsilon_{i1t} \quad (13)$$

For i = 1, ..., n; t = 1, 2, 3 and j = 0, 1. Equation of utility (13) can be transformed in the difference of utility: $U_{it} = U_{i1t}–U_{i0t}$ and the corresponding model is:

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

With $V_{it} = (V_{i1t}–V_{i0t}) = \alpha_t+\beta_t X_i+\gamma_t Z_{it}$; $Z_{it} = (Z_{i1t}–Z_{i0t})$; $\beta_t = \beta_{0t}-\beta_{1t}$; $\alpha_t = (\alpha_{0t}-\alpha_{1t})$. Data was generated on parameter value of α_t = -1, β_t = 0.5 and γ_t = 0.3. Structure of correlation/covariances to be examined are $r_{12} = \rho$ and $r_{13} = r_{32} = 0$:

$$\text{Cor}(U_i) = \begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Utility for t = 1 is in correlation with the utility for t = 2 and the correlation order ρ = 0, 0.1, 0.2,..., 0.9. The value of observation variable X_i and Z_{ijt} taken from normal distribution is:

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

Generated data with the assumption of ε_{ijt} is extreme value distributed and $\delta_i \sim N(0, \sigma^2)$. Estimation towards parameter was conducted by utilizing GEE and MLE for mixed logit model.

In order to estimate parameter using MLE, double integration was required and for this purpose, Halton sequence can be implemented. Simulation program for Halton can be adopted from library randtoolbox. For optimizing log-likelihood function, optim function by available BFGS method in R.2.10 program was used (R Development Core Team, 2009).

RESULT OF SIMULATION

Figure 1-21 depicted the effect of correlation value to the each estimator bias. From results, it can be concluded that if assumption by using GEE does not appropriate, the utilities that are correlated each other will be bias. Utility 1 (U_{i1}) is in correlation with utility 2 (U_{i2}) and both are not correlated to utility 3 (U_{i3}). Therefore, value of correlation only influences on parameters within U_{i1} and U_{i2}. On both utilities, the estimator bias appreciable and comparable respect to the correlation values (Fig. 1-9).

- Estimation by using MLE is more accurate compared to as in using GEE. At bigger correlation between utilities, estimator on GEE will has higher deviation (Fig. 1-9) and this is caused misassumption. Model Logit model estimated by using GEE is based on the assumption that error has variance of $\pi^2/3$ while for data, the variance is $(\sigma^2+\pi^2/3)$. As the assumption fulfilled, Mixed Logit model can estimate all parameter appropriately (Fig. 11-20)

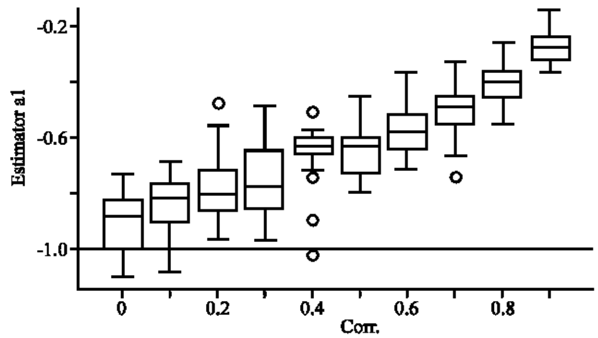

Fig. 1: Box plot α_1 (t = 1) in GEE

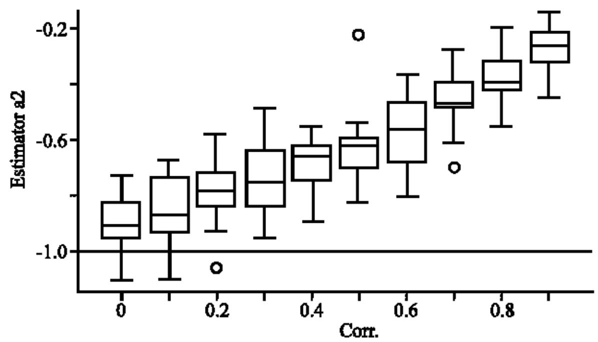

Fig. 2: Box plot α_2 (t = 2) in GEE

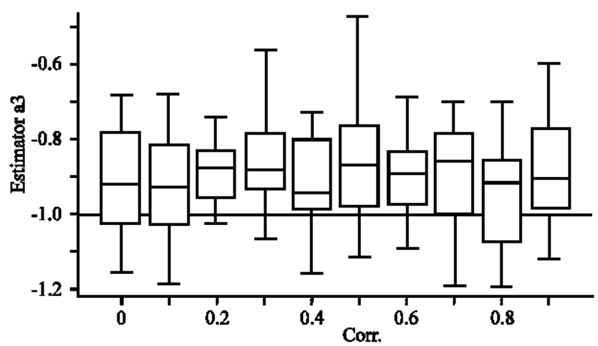

Fig. 3: Box plot α_3 (t = 3) in GEE

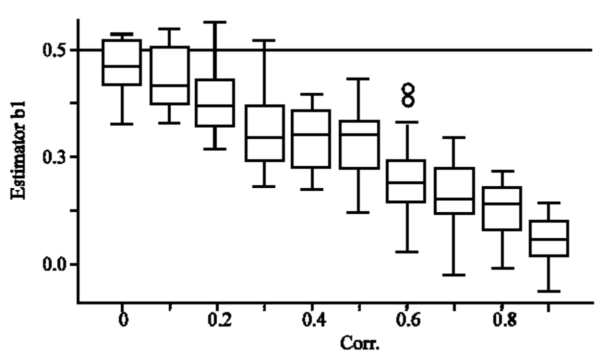

Fig. 4: Box plot $\boldsymbol{\beta}_1$ (t = 1) in GEE

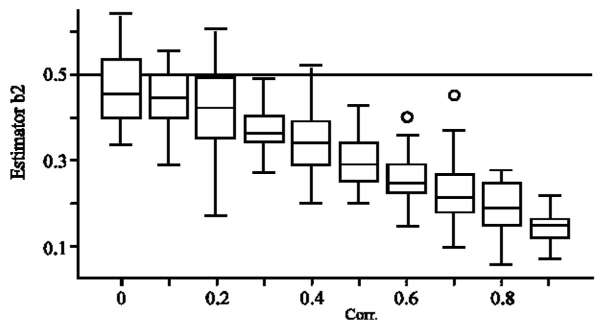

Fig. 5: Box plot $\boldsymbol{\beta}_2$ (t = 2) in GEE

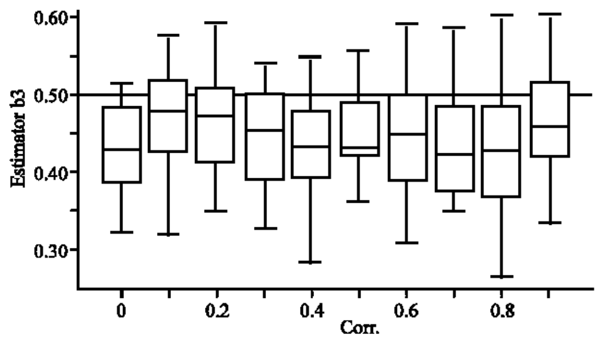

Fig. 6: Box plot $\boldsymbol{\beta}_3$ (t = 3) in GEE

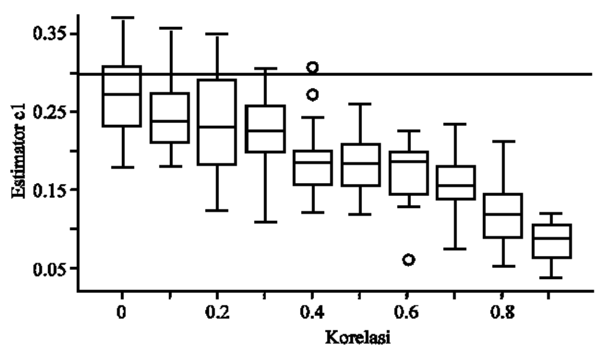

Fig. 7: Box plot γ_1 (t = 1) in GEE

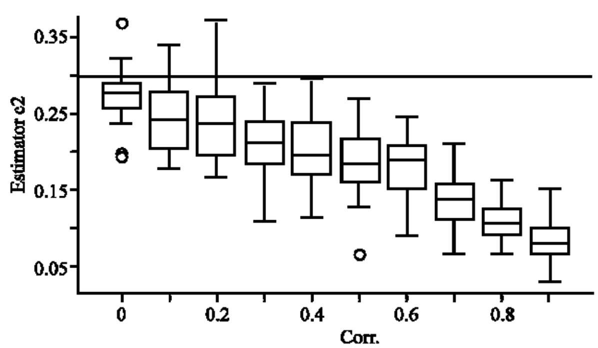

Fig. 8: Box plot γ_2 (t = 2) in GEE

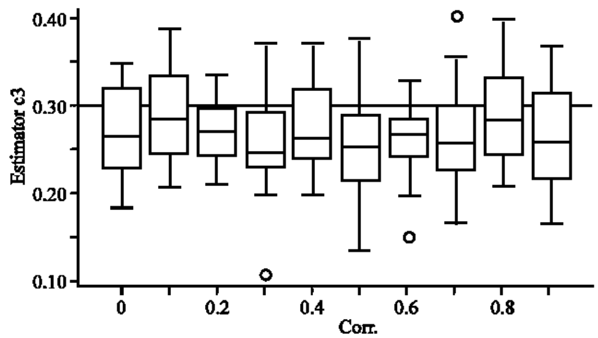

Fig. 9: Box plot γ_3 (t = 3) in GEE

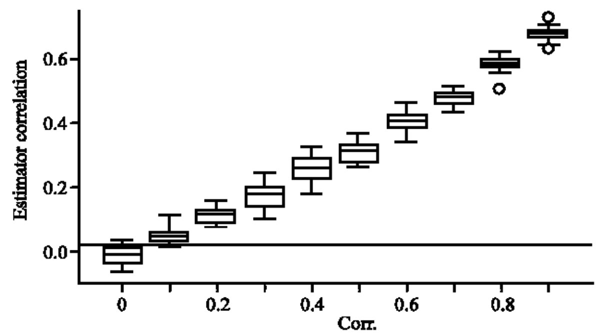

Fig. 10: Box plot ρ in GEE

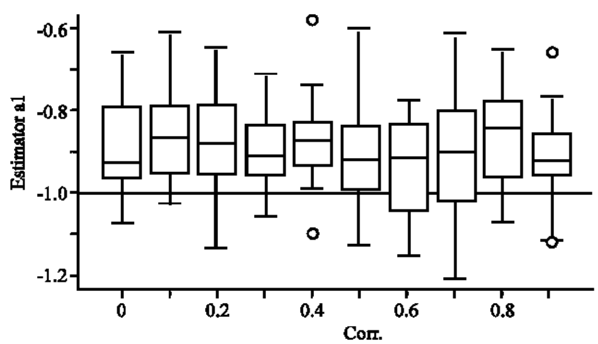

Fig. 11: Box plot α_1 (t = 1) in MLE

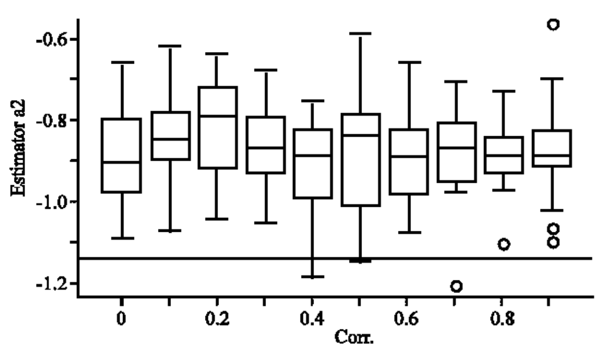

Fig. 12: Box plot α_2 (t = 2) in MLE

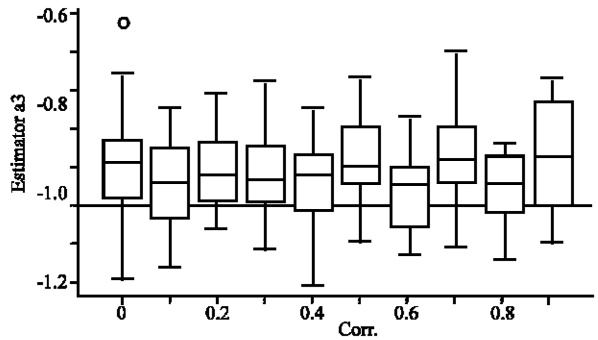

Fig. 13: Box plot α_3 (t = 3) in MLE

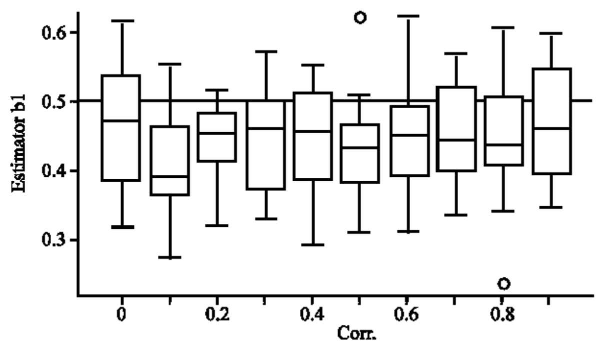

Fig. 14: Box plot β_1 (t = 1) in MLE

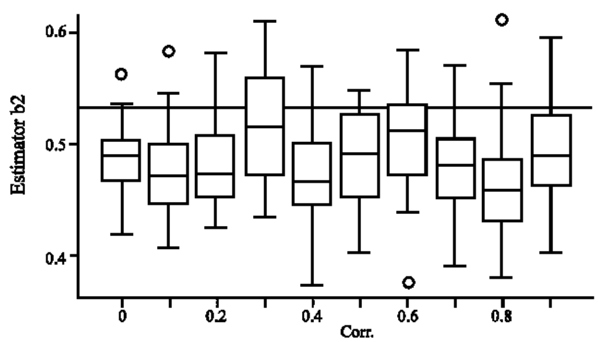

Fig. 15: Box plot β_2 (t = 2) in MLE

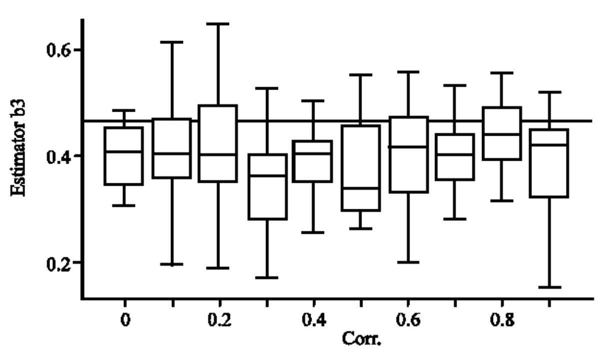

Fig. 16: Box plot β_3 (t = 3) in MLE

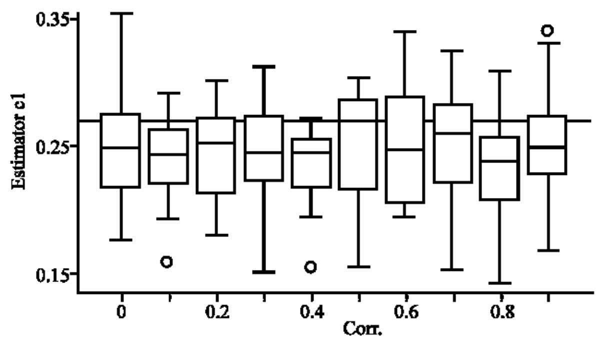

Fig. 17: Box plot γ_1 (t = 1) in MLE

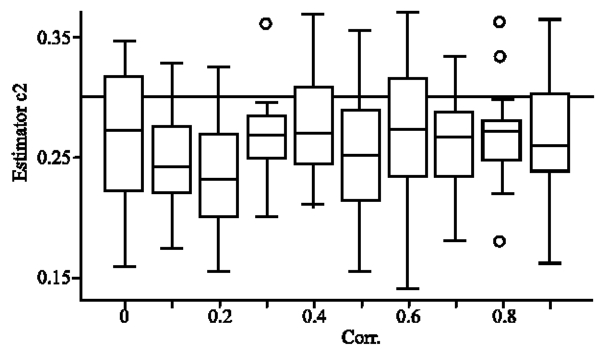

Fig. 18: Box plot γ_2 (t = 2) in MLE

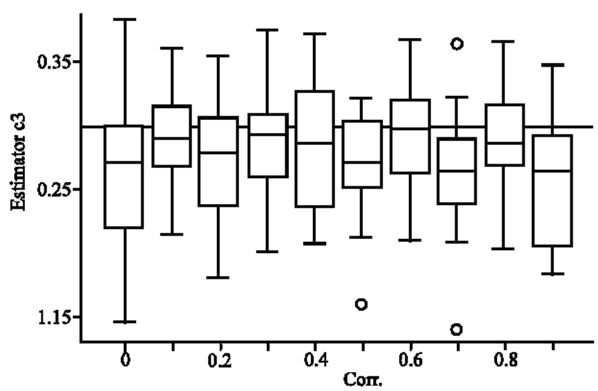

Fig. 19: Box plot γ_3 (t = 3) in MLE

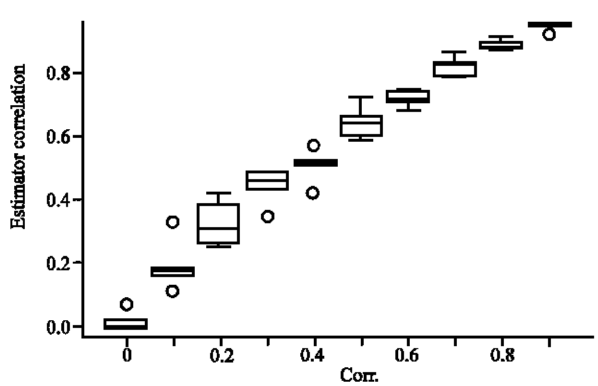

Fig. 20: Box plot ρ in MLE

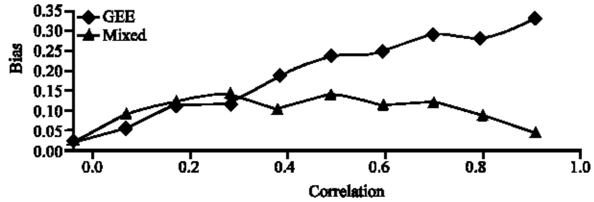

Fig. 21: Graph of a bias ρ in GEE and MLE

- Correlation parameter estimator for MLE is better (smallet bias) compared to as for GEE (Fig. 20, 21) more over, Correlation parameter estimator for GEE tends to be underestimate (Fig. 10)
- In low correlation (less than 0.5), MLE is robust respect to assumption deviation and can estimate precisely. Deviation of assumption in GEE method produce in bias estimator

CASE STUDY

Data are obtained from library Ecdat in R.2.10.0 which are data Heating. Data are resulted from survey of room heater household in California. From 850 families, there are several variables observed. Y_i is heater system utilized, that are gc (gas central), gr (gas room), ec (electric central) and er (electric room). X_{1i} is income of each family (income). X_{2i} is age of familyhed (Agehed). X_{3i} is amount of room in house (rooms) where, I = 1, ..., 850. Furthermore, multivariate approach with two responses was utilized. Y_{i1} is source of energy (g = 1) and electricity (e = 0). Y_{i2} is application type, central controled (c = 1) and room controled (r = 0). Characteristic variable for each response is defined in Table 1. From those data, tabulated data between energy source (Y_1) and application type (Y_2) (Table 2).

Based in Chi-squared Pearson's test, obtained value $\chi^2 = 95.8792$ (p-value<2.2e-16), means that energy source choosen by respondents and application type are not independent.

If there is correlation among responses, parameter estimator will be bias. On this condition, Mixed Logit model is more accurate in estimating parameter compared with as in discrete choice model. For three approximations used; univariate, GEE, mixed logit model, instalation cost and operational cost are the significant variables influence to the Y variable. In contrast, effect of individual charactersitics of X_{1i}, X_{2i} and X_{3i} to the response is not significant.

Equation of marginal probability on independent model and GEE for energy source is as follow:

$$P(Y_{i1}=g)=\frac{\exp(\gamma_{1S}(Z_{1ig}-Z_{1ie})+\gamma_{2S}(Z_{2ig}-Z_{2ie}))}{\exp(\gamma_{1S}(Z_{1ig}-Z_{1ie})+\gamma_{2S}(Z_{2ig}-Z_{2ie}))+1},\ P(Y_{i1}=e)=1-P(Y_{i1}=g)$$

and the probability equation for application type is:

$$P(Y_{i2}=c)=\frac{\exp(\gamma_{1T}(Z_{1ic}-Z_{1ir})+\gamma_{2T}(Z_{2ic}-Z_{2ir}))}{\exp(\gamma_{1T}(Z_{1ic}-Z_{1ir})+\gamma_{2T}(Z_{2ic}-Z_{2ir}))+1},\ P(Y_{i2}=r)=1-P(Y_{i2}=c)$$

Estimator value for each parameter is presented in Table 3.

Table 1: Characteristic variabel of election

Response	Instalation cost	Operational cost
$\mathbf{Y_{i1}}$		
g	Z_{1ig} = icS.1	Z_{2ig} = ocS.1
e	Z_{1ie} = isS.0	Z_{2ie} = ocS.0
$\mathbf{Y_{i2}}$		
c	Z_{1ic} = icT.1	Z_{2ic} = ocT.1
r	Z_{1ir} = icT.0	Z_{2ir} = ocT.0

Table 2: Crossed tabulation of energy source and application type

	Source	
Type	g	e
c	573	64
r	129	84

Table 3: Estimator for three models

Parameter	Independent	GEE	Mixed Logit
γ_{1S}	-0.003134484 (6.118290)	-0.002136034 (3.671803)	-0.0042593669 (5.357086)
γ_{1S}	-0.004997236 (182.300896)	-0.004964913 (196.656300)	-0.0034425002 (6.214808)
γ_{1T}	-0.004902223 (56.478941)	-0.004620066 (54.791350)	-0.0041680429 (39.067627)
γ_{1T}	0.006265215 (6.848393)	0.006951781 (9.265083)	-0.0008583833 (2.650140)
ρ	-	0.3131641 (47.84975)	0.6280863 (20.944647)

Values in blanket is statistic Wald

Equation of marginal probability for heater system on Logit model is:

$$P(Y_{i1}=g)=\int_{-\infty}^{\infty} l_{iS}(Z_g, Z_e \mid \xi_i)\,\varphi(\xi_i)\,d\xi_i \text{ dan } P(Y_{i1}=e)=1-P(Y_{i1}=g)$$

with:

$$l_{iS}(Z_g, Z_e \mid \xi_i)=\frac{\exp(\gamma_{1S}(Z_{1ig}-Z_{1ie})+\gamma_{2S}(Z_{2ig}-Z_{2ie})+\sigma\xi_i)}{\exp(\gamma_{1S}(Z_{1ig}-Z_{1ie})+\gamma_{2S}(Z_{2ig}-Z_{2ie})+\sigma\xi_i)+1}$$

ξ_i has normal distribution with density ϕ (ξ_i) and:

$$\sigma^2=\frac{\rho\pi^2/3}{1-\rho}$$

Equation of marginal probability for the variable of application type is:

$$P(Y_{i1}=c)=\int_{-\infty}^{\infty} l_{iT}(Z_c, Z_r \mid \xi_i)\,\varphi(\xi_i)\,d\xi_i \text{ dan } P(Y_{i2}=r)=1-P(Y_{i2}=c)$$

With:

$$l_{iT}(Z_c, Z_r \mid \xi_i)=\frac{\exp(\gamma_{1T}(Z_{1ic}-Z_{1ir})+\gamma_{2T}(Z_{2ic}-Z_{2ir})+\sigma\xi_i)}{\exp(\gamma_{1T}(Z_{1ic}-Z_{1ir})+\gamma_{2T}(Z_{2ic}-Z_{2ir})+\sigma\xi_i)+1}$$

Equation of combined probability on mixed logit model is:

$$P(Y_{i1}, Y_{i2}) = \int_{-\infty}^{\infty} (l_{iS})^{y_{i1}} (1 - l_{iS})^{(1-y_{i1})} (l_{iT})^{y_{i2}} (1 - l_{iT})^{(1-y_{i2})} \varphi(\xi_i)\, d\xi_i$$

If $Y_{i1} = g$, so $y_{i1} = 1$ and if $Y_{i1} = e$ so $y_{i1} = 0$. As $Y_{i2} = c$, $y_{i2} = 1$ and as $Y_{i2} = r$ $y_{i2} = 0$.

By using mixed logit model, correlation value between energy source and application source is 0.628 otherwise by using GEE correlation value is 0.313. From simulation result, it is noted that corelation estimator for GEE tend to be underestimate, therefore, estimator correlation on Mixed Logit model is more trusted.

CONCLUSION

Mixed Logit model is based on assumption that error component has extreme value distribution and parameter has normal distribution. Based on data simulation, if there is a violation to the assumption (specifically related to variance) GEE estimator will be bias. Estimator bias will be higher at higher variance value. Estimator bias in GEE can be reduced by MLE method. The correlation of parameter estimator in GEE tends to be underestimate but MLE can be used for estimating correlation between parameters accurately.

REFERENCES

Akanda, M.A.S., M. Khanam and M.A. Islam, 2005. Comparison of goodness-of-fit tests for GEE modeling with binary responses to diabetes mellitus. J. Applied Sci., 5: 1214-1218.

Alpu, O. and H. Fidan, 2004. Sequential probit model for infant mortality modelling in Turkey. J. Applied Sci., 4: 590-595.

Antoine, R., A. Sahai and P. Chami, 2007. Meta analysis of panel data generated by a set of randomized controlled trials. J. Applied Sci., 7: 951-957.

Chaganty, N.R. and H. Joe, 2004. Efficiency of generalized estimating equations for binary responses. J. Royal Stat. Soc. Ser. B (Stat. Methodol.), 66: 851-860.

Chong, E.K.P. and S.L. Zak, 1996. An Introduction to Optimization. John Wiley and Sons, Inc., New York.

Damisa, M.A., R. Samndi and M. Yohanna, 2007. Women participation in agricultural production: A probit analysis. J. Applied Sci., 7: 412-416.

Diggle, P., K.Y. Liang and S.L. Zeger, 1996. Analysis of Longitudinal Data. Oxford University Press, New York.

Geweke, J.F., M.P. Keane and D.E. Runkle, 1997. Statistical inference in the multinomial multiperiode probit model. J. Econometrics, 80: 125-165.

Greene, W., 2005. Econometric Analysis. 5th Edn., Prentice Hall, Englewood Cliffs, New Jersey.

Hajivassiliou, V., D. McFadden and P. Ruud, 1996. Simulation of multivariate normal rectangle probabilities and their derivatives: Theoretical and computational results. J. Econ., 72: 85-134.

Lechner, M., S. Lollivier and T. Magnac, 2005. Parametric binary choice models. Discution Paper No. 2005-23, University of St. Gallen. http://ideas.repec.org/p/usg/dp2005/2005-23.html.

Liang, K.Y. and S.L. Zeger, 1986. Longitudinal data analysis using generalized linear models. Biometrika, 73: 13-22.

Lipsitz, S.R., N.M. Laird and D.P. Harrington, 1990. Maximum likelihood regression methods for paired binary data. Stat. Med., 9: 1517-1525.

Munizaga, M.A. and R. Alvarez-Daziano, 2004. Testing mixed logit and probit models by simulation. TRB 2005 Annual Meeting CD-ROM.

Nugraha, J., 2000. Estimating parameter on bivariate binary logistic regression. Proceeding of National Seminar, Faculty of Mathematic and Natural Science, UGM Yogyakarta.

Nugraha, J., S. Guritno and S. Haryatmi, 2006. Logistic regression model for multivariate binary response using generalized estimating equation. Proceeding of National Seminar MIPA, UNY Yogyakarta.

Nugraha, J., S. Guritno and S. Haryatmi, 2008. Likelihood function and its derivatives of probit model on multivariate binary response. J. Kalam, 1: 28-35.

Nugraha, J., S. Guritno and S. Haryatmi, 2009. Comparison of logit model and probit model on multivariate binary response. International Conference on Natural and Material Sciences (NAMES 2009), Faculty of Mathematic and Natural Science -UNLAM Banjarmasin.

Nugraha, J., S. Guritno and S. Haryatmi, 2009. Estimating of parameter of logit model on multivariate binary response using MLE and GEE. J. Ilmu Dasar, 10: 82-92.

Prentice, R.L., 1988. Correlated binary regression with covariates specific to each binary observation. Biometrics, 44: 1033-1048.

R Development Core Team, 2009. R: A Language and Environment Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Sarkar, S.K. and H. Midi, 2010. Importance of assessing the model adequacy of binary logistic regression. J. Applied Sci., 10: 479-486.

Sarkar, S.K., H. Midi and S. Rana, 2011. Detection of outliers and influential observations in binary logistic regression: An empirical study. J. Applied Sci., 11: 26-35.

Train, K.E., 2003. Discrete Choice Methods with Simulation. Cambridge University Press, UK., ISBN-10: 0521017157.

Zadkarami, M.R., 2008. Bootstrapping: A nonparametric approach to identify the effect of sparsity of data in the binary regression models. J. Applied Sci., 8: 2991-2997.