Dhaka Univ. J. Sci. 59(1): 65-70, 2011 (January)

Extended Mallow's C_p in GEE: An Application to Maternal Morbidity Data

Rozana Rahman[1], Md. Anower Hossain[1] and M. Zakir Hossain[2]

[1]Institute of Statistical Research and Training (ISRT), Dhaka University, Dhaka-1000, Bangladesh

[2]Department of Statistics, Biostatistics and Informatics, Dhaka University, Dhaka-1000, Bangladesh

Received on 18. 12. 2009. Accepted for Publication on 28. 08. 2010

Abstract

Model selection is a vital issue as models are used for purposes of interpretation or prediction. While there are varieties of model selection criteria, fewer options for the longitudinal data analysis using Generalized Estimating Equations (GEE). Recently, an extended version of Mallow's C_p (GC_p) is suggested as model selection criterion which can be used for a marginal longitudinal model like GEE. In this study, an application of GC_p is demonstrated to select underlying model with important covariates associated with pregnancy complications for Bangladeshi women. Further statistical inferences are drawn for selected model by GEE.

Key words: Extended Mallow's C_p , GEE and Maternal Morbidity.

I. Introduction

For last few decades one of the most concerning topic was to reduce maternal mortality, morbidity and improving maternal health. However, women of developing countries like Bangladesh are still suffering from different life threatening diseases related to pregnancy during antenatal period and lost their life during pregnancy and post-partum period. Community level data on maternal morbidity in developing countries are inadequate as most studies are based on data collected from clinic or hospitals. But a huge segment of women still does not seek for such facilities; available figures do not represent the actual nature and magnitude of this problem. To present the actual enormity of the problem, Bangladesh Institute of Research for Promotion of Essential and Reproductive Health and Technologies (BIRPERHT) conducted a prospective survey on maternal morbidity in Bangladesh. In this study, a number of selected women were followed-up during their pregnancy and post-partum period and key variables related to pregnancy as well as presence or absence of any complication during this time are documented over the follow-up period for each of the selected women.

Several measurements were taken from each woman as they were followed on regular basis throughout the pregnancy which constituted a repeated measured data, often termed as longitudinal data. Longitudinal studies typically have two advantages which are increased power and robustness to model selection, suggested by Liang and Zeger1 (1992). Longitudinal data sets are comprised of repeated observations of an outcome variable and a set of covariates for each of many subjects. As repeated observations are made on each subject, correlation is anticipated among a subjects' measurement. A goodness of fit explains whether explanatory variables have significant effects to the response variable or not. But the correlations among the response variable have some influence over the fit. So analyzing the data from the repeated measures studies, one must account for this autocorrelation to make correct inference. Statisticians developed different methods for analyzing longitudinal data in the field of survival analysis for estimating the survival function, identifying the risk and prognostic factors of a particular disease, defining the relationship of the risk factors with the disease variable over a period of time. The outcome variables of longitudinal data may arise as continuous, categorical or count. In our present study we considered the longitudinal data with binary responses of subject.

Generalized Estimating Equations (GEE), proposed by Liang and Zeger2 (1986), has become an important strategy in the analysis of correlated data which may arise from longitudinal studies. GEE can be used for analyzing both continuous and discrete multivariate responses with in the generalized linear model frame work. It does not require the complete specification of the joint distribution of the repeated measurements. The GEE approach is an extension of quasi-likelihood to longitudinal data analysis. The main idea behind the quasi-likelihood method is to avoid a fully specified distribution for the response variable, when one is uncertain about the random mechanism by which the data were generated.

The GEE approach is based on the first two moments of the outcome variables under the assumption that variance is a known function of the mean. The method avoids the need for multivariate distributions by only assuming a functional form for the marginal distribution at each time. This method can provide consistent estimators of the regression parameters if the specification of the marginal means is correct. Since the true correlation among the repeated responses is unknown, GEE offers to take a working correlation for analysis. Among them, some common specifications are independence, exchangeable correlation, autoregressive correlation, pairwise correlation etc (Fitzmaurice et. al.[3] 1993). Correctness of the specification of the working correlation is not so important in a GEE analysis because the resulting regression coefficient estimators are still consistent even when the working correlation structure is miss-specified to some extent (Liang and Zeger[2], 1986). By choosing the correlation structure closer to the true correlation the efficiency of the estimates can be increased. Another advantage of GEE method is that it gives the robust estimates of the variance.

On the other hand, model selection is the task of choosing a model with the correct inductive bias, which in practice means selecting parameters in an attempt to create a model of optimal complexity for the given data (Sewell4, 2006). For non-likelihood based methods e.g. GEE, some model selection criteria have recently been developed. Some methods based on Bootstrapping and Cross validation are established (Cantoni, Field, Flemming, Ronchetti5, 2007; Pan6 2001 and Pan and Le7 2001). A modified Akaike's Information Criterion (mAIC), which is based on the quasi-likelihood function, was proposed as a model selection criterion for GEE (Pan8, 2001). Cantoni et al.[9] (2005) suggested a generalized version of Mallow's $C_P(GC_p)$ suitable for use with both parametric and non-parametric models that provides an estimate of a measure of model's adequacy for prediction. He also derived the form of GC_p for a marginal longitudinal model.

In the study, we used Mallow's $C_P(GC_p)$ in GEE to select best underlying model from a given set of covariates for maternal morbidity data. Also, we fitted GEE for the selected model and attempt to identify the significant risk factors with their magnitude which are associated with major pregnancy related complications during antenatal period.

II. Data and Variables

The study used data from the survey on maternal morbidity in Bangladesh. The survey was conducted from November 1992 to December 1993 by the Bangladesh Institute for the Research for promotion of Essential and Reproductive Health Technologies (BIRPERHT). The data were collected using both cross-sectional and prospective study designs. This study was based on the data from the prospective component of the survey. A number of papers have been published using this data set, e.g., Islam et al.10 (2004), Gulshan et al.11 (2005), Chakraborty et al.12 (2003),and Latif et al.13 (2008).

A multistage sampling design was used for collecting the data for this study. Districts were selected randomly in the first stage, one district from each division. Then thanas were selected randomly in the second stage, one thana from each of the selected districts. At the third stage, two unions were selected randomly from each selected thana. The subjects comprised of pregnant women with less than 6 months duration in the selected unions. All the selected pregnant women from the selected Unions were followed on regular basis (roughly at an interval of 1 month) throughout the pregnancy. Again, the subjects were followed at the time of delivery for a full term pregnancy and 90 days after delivery or 90 days after any other pregnancy outcome. A total of 1020 pregnant women were interviewed in the follow-up component of the study. The survey collected information on socio-economic and demographic characteristics, pregnancy-related care and practice, morbidity during the period of follow-up as well as in the past, information concerning complications at the time of delivery and during the postpartum period, etc. Here the number of follow-ups for each individual is not equal. In the study, the data of first four consecutive antenatal visits are considered and we have 549 such women's information for the analysis. This study makes an attempt to identify the risk factors associated with maternal morbidity in the antenatal period. To identify the morbid cases during pregnancy period we have considered at least one of the life-threatening complications which are-Hemorrhage, Edema, Excessive vomiting and Fits or Convulsion. The response variable can be defined below

$$Y = \begin{cases} 1, & \text{if the women suffers from at least} \\ & \text{one of the major complication} \\ 0, & \text{Otherwise} \end{cases}$$

Though data contain different factors related to pregnancy, here we have considered six factors as covariates for the purpose of our study, which are level of education of the respondents (EL), age at marriage (AM), economic status of the respondents (ES), gainful employment (GE), wanted pregnancy (WP) and Food Supplement (FS). All covariates are converted to the binary variables with two categories 0 and 1 where '0' is taken as a reference category implies that the respondents having any formal schooling, age at marriage less than or equal 15years, less than average for economic status, not involved in any gainful employment, not desired pregnancy and not taken any special food supplement.

III. Methods

Generalized Estimating Equations

Generalized estimating equation (GEE) method is non-likelihood based method and is widely used in analysis of correlated outcomes. The GEE model for the correlated outcomes is defined by the first two marginal moments of the outcomes and the working correlation within individuals. The GEE estimators of the parameters in the mean model are consistent as long as the marginal means of the outcome are correctly specified. In addition, correct specification of the correlation individuals can improve the efficiency of estimators (Liang and Zeger[2], 1986; Pretice and Zhao[14], 1991; Paik[15], 1992). It is an extension of quasi-likelihood to longitudinal data analysis.

We suppose that there are n individuals in the study. Each individual is observed at T_i occasions, i = 1,2,…,n. For simplicity let each individual is observed for equal number of repetitions, say T occasions. So our description of GEE inclined to equal repetitions. Thus, we have a $T \times 1$ random vector of response for the i^{th} individual as $y_i = (y_{i1}, y_{i2}, \cdots, y_{iT})$, where the response variable y_{ij} is dichotomous. Let $X_{ij} = \left(X_{ij1}, X_{ij2}, \ldots\ldots\ldots\ldots, X_{ijk}\right)'$ be the vector of covariates corresponding to the j^{th} response of

the i^{th} subject, $x_{ij1}=1$ for all i, j. let us assume that y_{ij} follows a distribution from exponential family and the mean response vector is, $\mu_{ij}=\Pr\left(y_{ij}=1\middle|X_{ij}\right)$, j = 1,2,…,T, i=1,2,..,n; the covariate set can be expressed by the link function $h(.)$ as $\mu_{ij}=h^{-1}\left(\beta' x_{ij}\right)$, where $\beta=\left(\beta_1,\beta_2,.....,\beta_k\right)'$, vector of parameters. The probability of developing the disease for i^{th} individual at j^{th} occasion is μ_{ij} and not developing the disease is $1-p_{ij}=1-\mu_{ij}$. So the variance of y_{ij} is $p_{ij}\left(1-p_{ij}\right)=\mu_{ij}\left(1-\mu_{ij}\right)$. In addition, to the mean and covariance of responses Liang and Zeger[2] (1986) suggested taking $T\times T$ working correlation matrix for each y_i, denoted by $R_i(\alpha)$. Considering the following quasi-likelihood approach, Liang and Zeger developed the GEE for β of the form

$$U(\beta)=\sum_{i=1}^{n} D_i' V_i^{-1}\left(Y_i-\mu_i\right)=0$$

where, $D_i=\frac{\partial \mu_i}{\partial \beta}$ and V_i is a working or approximate covariance matrix of y_i, chosen by the investigator. This working covariance matrix can be expressed in the following form $V_i=A_i^{\frac{1}{2}} R_i(\alpha) A_i^{\frac{1}{2}}$,

where, $A_i=diag\left\{\text{var}\left(Y_{i1}\right),.....\text{var}\left(Y_{iT}\right)\right\}$, is a $T\times T$ diagonal matrix and $V\left(Y_{ij}\right)=\phi V\left(\mu_{ij}\right)$ is a function of known mean function and dispersion parameter, ϕ.

This leads to the estimating equations (1) of the form

$$U(\beta)=\sum_{i=1}^{n} X_i' A_i V_i^{-1}\left(Y_i-\mu_i\right)=0$$

The GEE approach allows the time dependence to be specified in a variety of ways. The form of the working correlation parameters $\alpha=\left(\alpha_1,\alpha_2,.....,\alpha_q\right)'$. Working correlation structure in GEE can be chosen from following common forms:

i. Independence correlation: $R_i(\alpha)=Corr\left(y_i\right)=I_T$, where *I* is an identity matrix of order $T\times T$.
ii. Exchangeable Correlation: $R_i(\alpha)=Corr\left(Y_{ij},Y_{ik}\right)=\alpha$; $j\neq k$.
iii. Autoregressive Correlation: $R_i(\alpha)=Corr\left(Y_{ij},Y_{ik}\right)=\alpha^{|j-k|}$; $j\neq k$. For all $k>m$, $\alpha^{|j-k|}>\alpha^{|j-m|}$.
iv. Unstructured or Pairwise Correlation: $\left(R_i(\alpha)\right)_{ij}=Corr\left(Y_{ij},Y_{ik}\right)=\alpha_{jk}$; $j\neq k$,

where, $\alpha_{j,j+1}=\alpha_{j+1,j}$, j = 1,2,…,T

Mallow's Cp Variable Selection Criteria

In multiple regressions a response variable is usually expressed as a function of several 'independent' or predictor variables and an attempt is made to find out a subset of variables which is best to meet some specific objective. Whether explicitly stated or implicitly assumed, the underlying objective of the existing method is to minimize some expected discrepancy based on the error of prediction.

Here, we are considering Mallow's C_p variable selection criterion. For the C_p method (Mallows[16], 1973) it is assumed that the predictor variables are fixed and not random. For a p term subset model Mallow's C_p statistics is defined as

$$C_p=\frac{ESS_p}{\hat{\sigma}^2}+2p-n$$

Where, *ESSp* is the error sum of square for the subset model and $\hat{\sigma}^2$ is the unbiased estimate of σ^2 from the full model. Subsets of variables that produce C_p small or at least $C_P<p$ are the desirable subsets. The C_p statistic is an estimate of total error

$$\Gamma s=\sum_{i=1}^{n} MSE(\hat{y}_{i,p})/\sigma^2$$

where, $\hat{y}_{i,p}$ denotes the ith fitted value for the subset model. The C_p statistic, therefore, measures the performance of the variables in terms of the standardized mean square error of prediction. It takes into account both the bias as well as the variance.

Extension of Mallow's C_p in GEE

Cantoni et al.[9] (2005) extened the Mallows's C_p (Mallows[16], 1973) criterion in an extent that it requires only the data and a model from which predicted values can be obtained. The technique can be applied to many different types of models, including those in which the classical assumptions, in particular the independence of variables and their normal distributions, do not hold. For example, binary

outcome data are not normally distributed. In addition, the independence of outcome variables does not hold when repeated measurements are taken on the same subject. Generalized Linear Models (McCullagh and Nelder[17], 1989) and Generalized Estimating equation (Liang and Zeger[2], 1986) allows us to model the data described above. In these cases, variable selection efforts mainly rely on the use of Wald-type tests. This can be unreliable, because, among other things, the choice of working dependence model can impact point estimates and significance levels.

Consider the general setting in which we have only observations y_i, i =1,…,K, and a model either parametric or non-parametric in form, from which we can obtain predicted values $\hat{y}_i$, i=1,……,K. We defined the rescaled weighted predictive squared error

$$\Gamma_p = \sum_{i=1}^{K} E\left[w_i^2 \left(\frac{y_i - \hat{y}_i}{\sigma v_i^{1/2}} \right) \left(\frac{\hat{y}_i - Ey_i}{\sigma v_i^{1/2}} \right)^2 \right],$$

where, $\hat{y}_i$ is the fitted value for sub model P and Ey_i and $V(y_i) = \sigma^2 v_i$ are the expected value and variance under the full model. The weight function may $w_i(.)$, can be defined the weighted sum of squared residuals by

$WSE = \sum_{i=1}^{K} w_i^2(r_i) r_i^2$, where, $r_i = \frac{y_i - \hat{y}_i}{\sigma v_i^{1/2}}$ and let

$\delta_i = \frac{\hat{y}_i - Ey_i}{\sigma v_i^{1/2}}$,

Then generalized version of Mallow's C_p is

$$GC_p = WSR - \sum_{i=1}^{K} E\left[w_i^2(r_i) \varepsilon_i^2 \right] + 2 \sum_{i=1}^{K} E\left[w_i^2(r_i) \varepsilon_i \delta_i \right],$$

where, $\varepsilon_i = \frac{y_i - Ey_i}{\sigma v_i^{1/2}}$.

The latter two terms comprise the correlation term necessary in order to make WSR unbiased for Γ_p.

Now, we consider the situation of Generalized Estimating Equations (Liang and Zeger[2], 1986). Under the usual regularity conditions for M-estimators (Huber[18], 1981) the estimator defined as the solution of Generalized Estimating Equations is asymptotically normally distributed with asymptotic variance $M^{1}QM^{1}$,

where , $M = \lim_{K \to \infty} \frac{1}{K} \sum_{i=1}^{K} D_i^T \Gamma_i^T V_i^{-1} \Gamma_i D_i$ and

$Q = \lim_{K \to \infty} \sum_{i=1}^{K} D_i^T V_i^{-1} Var(\psi_i) V_i^{-1} \Gamma_i D_i$. where,

$\Gamma_i = E\left(\widetilde{\psi}_i - \widetilde{c}_i \right)$ with $\widetilde{\psi}_i = \frac{\partial \psi_i}{\partial \mu_i}$ and $\widetilde{c}_i = \frac{\partial c_i}{\partial \mu_i}$.

Moreover,

$\psi_i = W_i(Y_i - \mu_i)$, $c_i = E(\psi_i)$, $W_i = W_i(X_i, y_i, \mu_i)$ is a diagonal $n_i \times n_i$ weight matrix containing weights w_{it} for t = *1,….,n_i*. Note that for classical GEE (Liang and Zeger[2], 1986), W_i should be taken as identity matrix and therefore $\Gamma_i = \mathrm{I}$ and $c_i = 0$. For such longitudinal models and writing $\psi(\varepsilon_{it}) = w(\varepsilon_{it}).\varepsilon_{it}$, GC_p from (6) becomes:

$$GC_p = WSR - \sum_{i=1}^{K} \sum_{t=1}^{n_i} E\left[\psi^2(\varepsilon_{it}) \right] + t_1 - t_2$$

where $t_1 \cong \frac{2}{\sigma K} \sum_{i=1}^{K} Tr\left[M^{-1} E\left(D_i^T Z_i a_i^T A_i^{-1} D_i \right) \right]$

with $Z_i = \Gamma_i^T V_i^{-1} (\psi_i - c_i)$, $a_i = (a_{i1}, \cdots, a_{in_i})^T$ and

$a_{it} = \psi(\varepsilon_{it}) \psi'(\varepsilon_{it})$ and

$t_2 \cong \frac{1}{\sigma^2 K^2} \sum_{i=1}^{K} Tr\left[E\left(B_i A_i^{-1} D_i M^{-1} \left(\sum_{j=1}^{K} D_j Z_j Z_j^T D_j \right) M^{-1} D_i^T A_i^{-1} \right) \right]$

with $B_i = diag\left(b_{i1} \ldots\ldots b_{in_i} \right)$ and

$bit = \psi(\varepsilon_{it}) \psi''(\varepsilon_{it}) - \psi^2(\varepsilon_{it}) / \varepsilon_{it}^2 + (\psi'(\varepsilon_{it}))^2$.

If the weights in (6) are chosen to be identically one, we obtain-

$$GC_p = \sum_{i=1}^{K} \sum_{t=1}^{n_i} r_{it}^2 - \sum_{i=1}^{K} n_i + 2 \sum_{i=1}^{K} \sum_{t=1}^{n_i} E\left[\varepsilon_{it} \delta_{it} \right],$$

where $E\left[\varepsilon_{it}^2 \right] = 1$, and where the term t_2 turned into exactly as zero. Then we obtain,

$$GC_p = \sum_{i=1}^{K} \sum_{t=1}^{n_i} r_{it}^2 - \sum_{i=1}^{K} n_i + \frac{2}{\sigma^2 K} \sum_{i=1}^{K} Tr\left(M^{-1} D_i^T A_i^{-2} D \right),$$

which can be obtained directly without simulation. Models with small values of CG_p will be preferred to others; detailed on Cantoni et al[9]. (2005).

IV. Results and Discussion

Selection of Best Models

The main objective of this paper is to show the application of GC_p in selecting the best model in GEE setup. With selected six covariates, there are 63 possible models for each correlation structure, all of them are examined and the best models with different number of covariates are shown in Table 1.

Among the six models taking all possible subsets with single covariate, the smallest value of classical GC_p is for Model I with covariate FS for all three correlation structures that considered in the analysis. Considering 15 models taking all possible pair of covariates, Model II, which includes WP and FS as covariates, is the best choice.

Among the twenty models with three covariates, the model with the covariates EL, WP and FS, is found to be the best one; we denote this model as Model III. The best model with four covariates (Model IV) includes the covariate GE in addition to the covariates of Model III. For five covariates, the best model includes the covariate AM in addition to the covariates of Model IV.

Table. 1. Best models with different number of covariates

Model	Number of covariates	Covariates contained by the model	GC_p		
			Independence	Exchangeable	Unstructured
I	1	FS	101.16	89.98	104.99
II	2	WP, FS	74.25	57.64	79.19
III	3	EL, WP, FS	74.61	51.6	78.62
IV	**4**	**EL, GE, WP, FS**	**73.56**	**40.57**	**73.51**
V	5	EL, AM, GE, WP, FS	88.57	46.56	83.06
VI	6	EL, AM, ES, GE, WP, FS	102.86	53.80	96.83

The only model with six covariates is denoted as Model VI which contains all the covariates that are considered in this study. Among all the six models (Model I up to Model VI), Model IV with covariates education level of respondents, gainful employment of the respondent, wanted pregnancy and food supplement , can be considered as the best model because the corresponding GC_p value is the smallest and this true for all correlation structures. For all cases, the selected best models are found to be the best model for all three considered correlation structures.

Analysis of Morbidity Data for the best Selected Model under Different Correlation Structures

The accompanying table shows the estimates of the parameters of the best model (Model IV) under considered three correlation structures, namely, independence, exchangeable and unstructured.

Table. 2. Estimates of the parameters of Model IV (with P-values in parenthesis)

Covariates	Independence		Exchangeable		Unstructured	
	Estimated Coefficient	Odds Ratio	Estimated Coefficient	Odds Ratio	Estimated Coefficient	Odds Ratio
Intercept	0.424 (0.000)	1.528	0.440 (0.000)	1.552	0.509 (0.000)	1.665
EL	-0.386 (0.000)	0.679	-0.392 (0.000)	0.675	-0.387 (0.000)	0.679
GE	-0.348 (0.000)	0.706	-0.407 (0.000)	0.665	-0.393 (0.000)	0.674
WP	-0.410 (0.000)	0.663	-0.386 (0.000)	0.679	-0.388 (0.000)	0.678
FS	-0.471 (0.000)	0.624	-0.483 (0.000)	0.616	-0.424 (0.000)	0.654

It is found that all the four covariates show negative association with the major morbid conditions during pregnancy period and are statistically significant irrespective of the choice of correlation structures. From the odds ratio, it is observed that the women with primary or higher education were less likely to develop any of the major pregnancy complications (hemorrhage, edema, excessive vomiting and fits/convulsion) during pregnancy period than those women with no schooling. The chance of occurring any of the major complications during pregnancy period is

less for women involved in gainful employment than those who did not involve. The women who desired the index pregnancy had lower risk of developing the pregnancy complications than who did not expect. The analysis shows that the probability of developing major complications during pregnancy period is less likely for women who took special food than those who did not take special food.

V. Discussion and Conclusion

In this study we used BIRPHERT data to show an application of recently proposed extended version of Mallow's $C_P(GC_p)$ as a model selection criterion. The GC_p serves great purpose in selecting underlying best model in situations, when the response is multivariate non-normal and full likelihood function is not specified. Among three correlation structures, namely, independence, exchangeable and unstructured, the GC_P values are computed for all possible subset models. This study also makes an attempt to identify the risk factors associated with major pregnancy complications during antenatal period and considered at least one of the major life-threatening pregnancy complications (hemorrhage, edema, excessive vomiting and fits or convulsion) as the response variable, where the six important covariates related with this study are considered, namely, education level of the respondents, age at marriage, economic status of the respondents, gainful employment, wanted pregnancy and food supplement.

It is found that the model with covariates, education level of respondents, gainful employment, wanted pregnancy and food supplement, is the best choice among all possible models under three different correlation structures in GEE. The women with primary or higher education are less likely to suffer from any of the major complications during pregnancy than those with no schooling. Previous studies show that low incidence of maternal morbidities was found among the educated women (Choolani and Ratnam[19], 1995). Chowdhury et al.[20] (2007) examined the trends in maternal mortality in Matlab, Bangladesh over 30 years and revealed female education and poverty reduction are two important variables in reducing the maternal mortality. The probability of developing the complications during pregnancy is less likely for the women with gainful employment. If the index pregnancy is desired, then it is more likely that the incidence of major complications would decline in antenatal period. In other words, an undesired pregnancy results the higher risk of complications during pregnancy period. Also the women who took special food during pregnancy were less likely to suffer any major pregnancy complications than who did not take.

Acknowledgement

We gratefully acknowledge the permission of the Director, BIRPERHT, in relation to use of data in this paper. The authors are indebted to the Ford Foundation for funding the data collection of the maternal morbidity study.

---

1. Liang KY, S.L. Zeger and B Qaqish, 1992. Multivariate regression analysis for categorical data. *Journal of Royal Statistical Society,* Series B, **54**: 3-40.
2. Liang, K.Y. and S.L. Zeger, 1986. Longitudinal data analysis using generalized linear models. *Biometrika,* **73** (1), 13-22.
3. Fitzmaurice, G.M. and N.M. Laird, 1993. A likelihood-based method for analyzing longitudinal binary responses. *Biometrika,* **80**(1), 141-151.
4. Sewell, M., 2006. Model selection. Research paper, Department of Computer Science University College London.
5. Cantoni, E., C. Field, J. M. FLemming and E. Ronchetti, 2007. Longitudinal variable selection by cross-validation in the case of many covariates. *Statistics in Medicine*, **26**(4), 919-930.
6. Pan, W. (2001). Model Section in Estimating Equations. *Biometrics*, **57, 2,** 529-234.
7. Pan, W. and C. T. Le, 2001. Bootstrap Model Selection in Generalized Linear Models. *Journal of Agricultural, Biological and Environmental Statistics,* **6,** 1, 49-61.
8. Pan, W., 2001. Akaike's Information Criterion in generalized Estimating Equations. *Biometrics*, **57**, 120-125.
9. Cantoni, E., J. M. FLemming, and E. Ronchetti, 2005. Variable selection for marginal longitudinal generalized linear models. *Biometrics*, **61,** 507-514 .
10. Islam, M. A., R. I. Chowdhury, N. Chakraborty, and W. Bari, 2004. A multistage model for maternal morbidity during antenatal, delivery and postpartum periods. *Statistics in Medicine*, **23**: 137-158.
11. Gulshan, J., R. I. Chowdhury, M. A. Islam, and H. H. Akhter, 2005. GEE models for maternal morbidity in rural Bangladesh. *Austrian Journal of Statistics*, **34**:295-304.
12. Chakraborty, N., M. A. Islam, R. I. Chowdhury, and W. Bari, 2003. Analysis of Ante-partum maternal morbidity in rural Bangladesh. Australian Journal of Rural Health, **11** : 22-27.
13. Latif, A.H.M., M. Z. Hossain, and M.A. Islam, 2008. Model Selection Using Modified Akaike's Information Criterion: An Application to Maternal Morbidity Data. *Austrian Journal of Statistics*, **37(2)**:175–184.
14. Pretice , R.L. and L.P. Zhao, 1991. Estimating equation for parameters in means and covariates of multivariate discrete and continuous responses. *Biometrics,* **47**, 825-839.
15. Paik, M.C. 1992. Quai-likelihood regression model with missing covariates, *Boimetrika*, **83,** 4, 825-834.
16. Mallows, C.L., 1973. Some comments on *Cp*. *Technometrics*, **15**, 661-675.
17. McCullagh, P. and J.A. Nelder, 1989. Generalized Linear Models, 2"d edition, Chapman and Hall, London.
18. Hurber, P. J., 1981. Robust Statistics, New York, Wiley.
19. Choolani, M. and S. S. Ratnam, 1995. Maternal morbidity: a global overview. *Journal of the Indian Medical Association*, **93**:36-40.
20. Chowdhury, M. E., R. Botlero, M. Koblinsky, S.K. Saha, G. Dieltiens, and C. Ronsmans, 2007. Determinants of reduction in maternal mortality in Matlab, Bangladesh: a 30-year cohort study. *Lancet*, 370: 1320-1328.