Dhaka Univ. J. Sci. 58(2): 219-223, 2010 (July)

Generalized Linear Mixed Models for Longitudinal Data Analysis: An Application to Maternal Morbidity Data

Muhammad Abu Shadeque Mullah, Nabila Parveen and M. Zakir Hossain

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

Received on 11. 08. 2009. Accepted for Publication on 29. 12. 2009

Abstract

This article discusses the application of Generalized Linear Mixed Models (GLMM) in which heterogeneity in regression parameters is explicitly modelled in order to analyze the longitudinal data related to the maternal morbidity. The most commonly used model selection criterion, Akaike's Information Criterion (AIC) has been used to select important covariates associated with the pregnancy related complications of Bangladeshi women. For testing the variance component in generalized linear mixed models, Likelihood Ratio Test (LRT) has been performed.

Key words: GLMM, AIC, LRT and Maternal Morbidity

I. Introduction

Although reproductive health care is a highly focused issue in the development of a country, millions of women experience life-threatening and other health related complications during pregnancy and post-partum period in developing countries like Bangladesh. In Bangladesh, around 16000 maternal deaths occurred due to pregnancy and delivery related problems in the year 2000 (Latif et al.[1], 2008). Poor access to services, low quality of care, high rate of maternal mortality and poor status of child health still remain as challenges of the health sector.

Despite pregnancy and delivery related complicacies are very common to Bangladeshi women, not too many attentions have been paid by the researchers in this issue. In recent times, a non-governmental organization, Bangladesh Institute of Research for Promotion of Essential and Reproductive Health Technologies (BIRPERHT) carried out a prospective survey on maternal morbidity in Bangladesh where the selected women were followed during the pregnancy and post-partum period. Along with important pregnancy-related variables, presence/absence of any complication during pregnancy is recorded over the follow-up period for each of the selected women.

As repeated measurements were made from each woman over different time points, the data can be phrased as longitudinal data. Longitudinal data are characterized by the fact that repeated observations for a subject tend to be correlated. This correlation presents additional prospects and challenges for analysis. There are two distinct approaches to longitudinal data analysis namely; subject-specific approach and population-averaged approach. In the "subject-specific" (SS) approach the heterogeneity across subjects can be explicitly modelled. The mixed model is an example where the subject-specific effects are assumed to follow a parametric distribution across the population (Zeger et al.[2], 1988). Mixed linear models (Laird and Ware[3], 1982; Ware[4], 1985) for continuous longitudinal data are in common use. Mixed generalized linear models for non-normal outcomes have recently become a research focus (See Stiratelli, Laird, and Ware[5], 1984; Anderson and Aitkin[6], 1985; and Gilmour, Anderson, and Rae[7], 1985 for applications to binomial data).

On the other hand, the population-averaged response can be modelled as a function of covariates without explicitly accounting for subject to subject heterogeneity. The regression coefficients have interpretation for the population rather than for any individual and hence the term "population-averaged"(PA) model (also known as marginal model) in this case. The principal distinction between SS and PA models is whether the regression coefficients describe an individual's or the average population response to changing covariates. SS models are desirable when the response for an individual rather than for the population is the focus.

In case of SS approach, models can be estimated by two distinct techniques: likelihood based and estimating equation based methods. The likelihood based methods require complete specification of the joint distribution of the multivariate responses, whereas the estimating equation based methods can be employed when joint distribution is not fully specified. Zeger et al.[2], 1988 described how generalized estimating equation (GEE) methodology, an estimating equation based method can be applied to estimate SS models.The present study considers the SS approach where model has been estimated via likelihood based method.

Model selection is an important task of data analysis which leads to select a "best" statistical model from a set of potential models, given data. That is selecting the best subset of the covariates from the available covariates in the data. Usually model selection is done by using a specific criterion. For likelihood-based methods, Akaike's Information Criterion (AIC) (Akaike[8], 1973) is widely used as a model selection criterion.

The main focus of the present study is to select the best models from a given set of covariates when the outcome is multivariate binary. The generalized linear mixed model is considered for modeling multivariate binary response and Akaike's information criterion is used to select the best subset of the available covariates. The test of homogeneity for the generalized linear model with random intercept for the best selected model has been performed by using likelihood ratio test applied to maternal morbidity data.

II. Data and Variables

The present study is based on the data from the maternal morbidity survey in Bangladesh conducted by the Bangladesh Institute of Research for Promotion of Essential and Reproductive Health Technologies (BIRPERHT) during the period of November 1992 to December 1993. A number of papers have been published using this data set (See Islam et al.[9], 2004; Gulshan et al.[10], 2005; Chakraborty et al.[11], 2003 and Latif et al.[1], 2008).

In the survey, a multistage sampling design was used where in the first stage the districts were randomly selected in such a way that exactly one district was chosen from each division. One thana was randomly selected from each of the chosen districts in the second stage and two unions were randomly selected from each of the selected thanas in the third stage. The sample comprised of all the pregnant women of duration at most six months from the selected unions. A total of 1020 selected pregnant women were followed till 90 days after delivery to collect information regarding socio-economic and demographic characteristics, pregnancy related care and practice, morbidity during the period of follow-up and in the past, complications at the time of delivery and during the postpartum period, etc.

With a view to identify the important factors associated with pregnancy related complications, the present study considered the first four consecutive antenatal follow-ups for 549 pregnant women. To identify the morbid cases in the pregnancy period we have considered at least one of the major life-threatening complications namely; haemorrhage, oedema, excessive vomiting, and fits or convulsion. Thus the response variable is considered as binary taking the value 1 if at least one of the complications was present. Notationally,

$$y = \begin{cases} 1, \text{ if the woman suffers from at least one of the major complications} \\ 0, \text{ otherwise.} \end{cases}$$

In the midst of the available covariates, only six important covariates are considered which are: educational level of the respondents (EDLV), age at marriage (AGEM), economic status (ECOST), desired the index pregnancy (DIP), food supplement (FSUP), and gainful employment (GEMP). All these covariates are coded as binary with the reference categories, never attended school for educational level, 15 years or less for age at marriage, less than average for economic status, no for desired index pregnancy, food supplement, and gainful employment respectively.

III. Methods

Generalized Linear Mixed Models

Let y_{it} be the binary response and x_{it}, a p × 1 vector of fixed covariates at time t for woman i, where

t = 1, ..., n_i and i = 1, ... ,n. If $y_i = (y_{i1}, y_{i2}, \ldots, y_{ini})$ is a vector of correlated responses from the i^{th} woman then to analyze such correlated data, random subject (woman) effects can be added into the regression model to account for the correlation of the data. The resulting model is a mixed model including the usual fixed effects for the regressors plus the random effects. Here we assume that the data for a single woman are independent observations from a distribution belonging to the exponential family, but that the regression coefficients can vary from woman to woman according to a random effects distribution, denoted by F. That is, conditional on the random effects, it is assumed that the responses for a single subject are independent observations from a distribution belonging to the exponential family. Let z_{it} be a q x 1 vector of covariates (typically a subset of x_{it}) associated with a q x 1 random effect, b_i, and let the conditional mean $u_{it} = E(y_{it}|b_i)$. Under the mixed GLM, the responses for subject i are assumed to satisfy

$$g(u_{it}) = x'_{it}\beta + z'_{it}b_i \ , \ \ \text{var}(y_{it} \mid b_i) = v(u_{it}).\phi$$

where b_i is an independent observation from a mixture distribution, F. The functions *g* and *v* are referred to as the "link" and "variance" functions, respectively.

A random-intercept model, which is the simplest mixed model, augments the linear predictor with a single random effect for subject *i*,

$$g(u_{it}) = x'_{it}\beta + b_i$$

where b_i is the random effect (one for each subject). These random effects represent the influence of subject *i* on his/her repeated observations that is not captured by the observed covariates. These are treated as random effects because the sampled subjects are thought to represent a population of subjects, and they are usually assumed to be distributed as $N(0, \sigma_b^2)$. The parameter σ_b^2 indicates the variance in the population distribution, and therefore the degree of heterogeneity of subjects. The objective of analysis is to estimate the fixed effects coefficients, β, parameters of F, and possibly the scale parameter, ϕ.

Dichotomous Outcomes

The mixed-effects logistic regression model is a common choice for analysis of multilevel dichotomous data and is arguably the most popular GLMM. In the GLMM context, this model utilizes the logit link, namely

$$g(u_{it}) = \text{logit}(u_{it}) = \log\left(\frac{u_{it}}{1-u_{it}}\right) = \eta_{it} = x'_{it}\beta + b_i$$

Here, the conditional expectation $u_{it} = E(y_{it}|b_i)$ equals $P(y_{it} = 1|b_i)$, namely, the conditional probability of a response given the random effects (and covariate values).

This model can also be written as

$$P(y_{it} = 1|b_i) = g^{-1}(\eta_{it}) = g^{-1}(x'_{it}\beta + b_i)$$

where the inverse link function $g^{-1}(\eta_{it})$ is the logistic cumulative distribution function (cdf), namely

$$g^{-1}(\eta_{it}) = [1 + \exp(-\eta_{it})]^{-1}$$

The probit model, which is based on the standard normal distribution, is often proposed as an alternative to the logistic model (Gibbons and Bock,[12] 1987). For the probit model, the normal cdf and pdf replace their logistic counterparts.

Model Specification

Since y_{it} is a binary response, taking values of 0 or 1, a logistic mixed effects model for y_{it} has been considered which is given by the following three-part specification:

1. Conditional on a single random effect b_i, the y_{it} are independent and have a Bernoulli (p_{it}) distribution,

with $\text{var}(y_{it} \mid b_i) = E(y_{it} \mid b_i)\{1 - E(y_{it} \mid b_i)\}$, (i.e., $\phi = 1$).

2. The conditional mean of y_{it} depend upon fixed and random effects via the following linear predictor: $\eta_{it} = x'_{it}\beta + z'_{it}b_i = x'_{it}\beta + b_i$,

where $z_{it} = 1$ for all i = 1,2,…,n(=549), and t = 1, …, 4, with

$$\log\left\{\frac{\Pr(y_{it}=1|b_i)}{\Pr(y_{it}=0|b_i)}\right\} = \eta_{it} = x'_{it}\beta + b_i \quad \text{…….(1)}$$

That is, the conditional mean of y_{it} is related to the linear predictor by a logit link function.

3. The single random effects b_i is assumed to have a univariate normal distribution, with zero mean and variance σ_b^2.

Estimation

With generalized linear mixed effects models the joint distributions of both the vector of responses and the vector of random effects are fully specified. As a result, we can base estimation and inference on the likelihood function. Given the three-part specification of a generalized linear mixed effects model, the joint probability for y_i and b_i can be expressed as:

$$f(y_i, b_i) = f(y_i|b_i)f(b_i) = f(y_{i1}|b_i)f(y_{i2}|b_i)...f(y_{ini}|b_i).f(b_i) \quad [\because (y_{ij}|b_i) \text{ are independent}]$$

The likelihood function is

$$L(\beta, \sigma_b^2) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \int f(y_i, b_i) db_i = \prod_{i=1}^{n} \int f(y_i|b_i) f(b_i) db_i = \prod_{i=1}^{n} \int \prod_{t=1}^{n_i} f(y_{it}|b_i) f(b_i) db_i$$

$$= \prod_{i=1}^{n} \int \prod_{t=1}^{n_i} \left\{p_{it}^{y_{it}}(1-p_{it})^{1-y_{it}}\right\} \left\{ 1/\sqrt{2\pi\sigma_b^2}\ \exp(-1/2\sigma_b^2) b_i^2 \right\} db_i$$

$$= \prod_{i=1}^{n} \int \prod_{t=1}^{n_i} \left\{\frac{\exp(x'_{it}\beta + b_i)}{1+\exp(x'_{it}\beta + b_i)}\right\}^{y_{it}} \left\{\frac{1}{1+\exp(x'_{it}\beta + b_i)}\right\}^{1-y_{it}} \left\{ 1/\sqrt{2\pi\sigma_b^2}\ \exp(-1/2\sigma_b^2) b_i^2 \right\} db_i$$

$$= \prod_{i=1}^{n} \int \prod_{t=1}^{n_i} \{\exp(x'_{it}\beta + b_i)\}^{y_{it}} \left\{\frac{1}{1+\exp(x'_{it}\beta + b_i)}\right\} \left\{ 1/\sqrt{2\pi\sigma_b^2}\ \exp(-1/2\sigma_b^2) b_i^2 \right\} db_i$$

$$= \prod_{i=1}^{n} \int \left\{\exp(\beta \sum_{t=1}^{n_i} y_{it}x'_{it} + y_i b_i)\right\} \prod_{t=1}^{n_i} \left\{\frac{1}{1+\exp(x'_{it}\beta + b_i)}\right\} \left\{ 1/\sqrt{2\pi\sigma_b^2}\ \exp(-1/2\sigma_b^2) b_i^2 \right\} db_i \quad \text{……..(2)}$$

The ML estimates of β and σ_b^2 are simply those values of β and σ_b^2 that maximize this likelihood function. However, the likelihood function (2) cannot be simplified or evaluated in closed form as this integral can not be evaluated in closed form and hence maximizing values cannot be expressed in closed form either. The integrals in the above equation are well suited for evaluation with the aid of Gauss-Hermite quadrature. As with any "automatic" numerical integration method, there are situations in which Gauss-Hermite quadrature for models like the logit-normal will give inaccurate results (McCulloch[13], 2003), generally having to do with the placement of the evaluation points. An improvement on simple Gauss-Hermite quadrature is adaptive quadrature, as exemplified in SAS Proc NLMIXED (SAS Institute, 2001) and Rabe-Hesketh et al.[14] (2002), in which the point of evaluation of the integral is "centered" in order to improve accuracy.

Akaike's Information Criterion

Akaike's Information Criterion (AIC) (Akaike[8], 1973) is very powerful and widely used model-selection criterion based on the likelihood and asymptotic properties of the maximum likelihood estimator (MLE). It was introduced as an approximately unbiased estimator of the expected Kullback-Leibler information of the fitted model. Suppose $D = \{(y_i, x'_{ij})\}$ be the data at hand, where y_i is the response vector and x'_{ij} is a set of covariates. Also suppose we have a candidate model M and the true model M* with log-likelihood functions $L(\beta; D)$ and $L(\beta^*; D)$ respectively, where β and β^* are the corresponding regression parameters. A well-known measure of separation between two models is given by the Kullback- Leibler information (Kullback and Leibler[15], 1951), also known as the cross entropy. The Kullback-Leibler information between M and M* is

$$\Delta_0(\beta, \beta^*) = E_{M^*}[-2L(\beta, D)],$$

where the expectation E_{M^*} is taken with respect to the true model M*. From a set of candidate models, we would like to choose the model with smallest $\Delta_0(\beta, \beta^*)$. However, in practice, both β and β^* are unknown and, as such, we

have to estimate $\Delta_0(\beta,\beta^*)$. AIC was motivated as an asymptotically unbiased estimator of $E_{M^*}[\Delta_0(\hat{\beta},\beta^*)]$, where $\hat{\beta}$ is the maximum likelihood estimator under any competing model and the expectation is taken over the random $\hat{\beta}$ .Notationally,

$$AIC = -2L(\hat{\beta};D) + 2p \ , \quad \ldots\ldots. (3)$$

where p is the dimension of β . Akaike proposed using AIC as a model-selection criterion by selecting a model that minimizes AIC as the "best" model.

The Likelihood Ratio Test for Variance Components in GLMMs

In many situations, we are interested in testing whether the between-subject variation is absent in the mixed effects model. This is equivalent to testing variance component equal to zero. That is, the hypothesis of interest is

$$H_0 : \sigma_b^2 = 0 \text{ against } H_1 : \sigma_b^2 > 0.$$

In a regular hypothesis testing setting, a likelihood ratio test is the most commonly used test due to its desirable properties and the fact that it is easy to construct.

The likelihood ratio test statistic is given by

$$G^2 = 2(l_1 - l_0) \ , \ldots\ldots. (4)$$

where $l_1 = \log L(\hat{\beta},\sigma_b^2)$ is the unconditional maximized log likelihood and $l_0 = \log L(\tilde{\beta},\sigma_b^2 = 0)$ is the maximized log likelihood under $H_0 : \sigma_b^2 = 0$. Here, G^2 follows chi-square distribution with 1 degree of freedom.

IV. Results and Discussion

Choice of Best Models

An important objective of this paper is to select the best model within mixed model setup using AIC.

All possible models that can be considered from the selected six covariates are examined and the best models with different number of covariates are shown in Table 1. Among the six models with one covariate, the model with DIP as the only covariate (Model I) is found to be the best one because the corresponding AIC value is the smallest. Among the 15 models with two covariates, Model II, which includes DIP and EDLV as covariates, is the best choice. For three covariates, the model with the covariates DIP, EDLV, and GEMP 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 FSUP in addition to the covariates of Model III. For five covariates, the best model includes the covariate AGEM in addition to the covariates of Model IV. 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 II can be considered as the best model because the corresponding AIC value is the smallest.

Table.1. Best models with different number of covariates

Model	Best Model With Number of Variables	Selected Covariates	AIC
I	1	DIP	744.3
II	2	EDLV ,DIP	743.2
III	3	EDLV, DIP , GEMP	745.1
IV	4	EDLV, FSUP, DIP ,GEMP	746.1
V	5	EDLV, AGEM, FSUP, DIP, GEMP	747.5
VI	6	EDLV, AGEM, ECOST, FSUP, DIP, GEMP	749.3

Analysis of Morbidity Data Using "Best" Selected Model

The ML estimates of the fixed effects and variance component of the best model (Model II) are presented in Table 2. These results indicate that the covariate DIP is highly significant and that women's unwillingness of being pregnant is harmful, increasing the woman-specific rates of having the pregnancy related complications during the pregnancy period. More specifically, the odds of experiencing pregnancy related complications is about 2 (or $1/e^{-.6162}$) times as high among the women who oppose to be pregnant than those who support.

The other variable of the best model, education level, is found to have a non-significant effect on pregnancy related complications. Note that the variance of b_i is correctly estimated as evidenced by the very small standard error. The estimate of the variance of b_i, $\hat{\sigma}_b^2 = 2.48$, indicates that there is between-woman variability in experiencing the pregnancy related complications.

These results are quite different from that obtained by Latif et al[1]. (2008) using the same data but generalized estimating equation approach. They found FSUP as the only significant covariate irrespective of the choice of the correlation structure under the GEE setup even though AGEM was found to be significant (at 10 percent) only for unstructured correlation structure.

Table. 2. Estimates of the parameters of Model II

Parameter	Estimate	Standard Error	P-value
Intercept	-3.5826	0.3338	<.0001
DIP	-0.6162	0.2966	0.0382
EDLV	-0.4144	0.3027	0.1715
$\hat{\sigma}_b^2$	2.48	0.7517	0.0010

Optimization Technique: Dual Quasi-Newton

Integration Method : Adaptive Gaussian Quadrature

The likelihood ratio test for variance components in generalized linear mixed model gives

$$G^2 = 2(l_1 - l_0) = 2[-368.1-(-383.15)] = 30.1$$

(P-value = 0.000).

This indicates that there is significant between-woman variability in experiencing the complications during pregnancy period and the use of generalized linear mixed model instead of generalized linear model to analyze the maternal morbidity data is appropriate.

V. Conclusion

Reproductive health care is a greatly focused issue in the development of a country. Although pregnancy and delivery related complicacies are very common to Bangladeshi women, not too many studies have been conducted to identify the important covariates associated with such pregnancy related problems. The present study was aimed at analysing the BIRPHERT data using the generalized linear mixed model to identify the important covariates associated with the maternal complications. Among the six covariates we have considered in this analysis, educational level of the respondents (EDLV) and desired the index pregnancy (DIP) are found to be the best subset of the covariates among all possible subsets of covariates. The analysis of the best model shows that woman's desireness regarding their pregnancy has significant impact on major complicacies during pregnancy. It is evident from the analysis that the probability of developing some major complications during pregnancy is smaller for women who wanted to have a baby than those who did not, and as such, woman's opinion regarding pregnancy should essentially take into consideration.

The covariate education level of the women has been selected in the best model though not found to have a strong significant effect on pregnancy related complications. Previous studies show that female education plays an important role in reducing maternal mortality, more specifically, a low incidence of maternal morbidities was found among the educated women (Choolani and Ratnam[16], 1995). Chowdhury et al.[17] (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. Age at marriage is also an important covariate for pregnancy related studies in developing countries.

The likelihood ratio test for variance components in generalized linear mixed model indicates that there is significant between-woman variability in experiencing the complications during pregnancy period and justifies the application of generalized linear mixed model to analyze the longitudinal data related to maternal morbidity.

The result of this study can be used in applied works in the relevant areas. Furthermore, as the study is based on a large number of data set, the findings can be used by Government and Non-government organizations for public policy formulations.

Acknowledgement

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

---

1. Latif, A.H.M.M., M.Z. Hossain and M.A. Islam. 2008. ModelSelection Using Modified Akaike's Information Criterion: An Application to Maternal Morbidity Data. Austrian Journal of Statistics, **37(2):**175–184.
2. Zeger, S.L., K.Y. Liang and P.S. Albert. 1988. Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics,* **44**:1049-1060.
3. Laird, N. M. and J. H. Ware. 1982. Random-effects models for longitudinal data. *Biometrics* 38: 963-974.
4. Ware, J. H. 1985. Linear models for the analysis of serial measurements in longitudinal studies. *American Statistician,* **39**:95-101.
5. Stiratelli, R., N. M. Laird and J. H. Ware. 1984. Random-effects models for serial observations with binary responses. *Biometrics*, **40**: 961-971.
6. Anderson, D. A. and M. Aitkin. 1985. Variance component models with binary response: Interviewer variability. *Journal of the Royal Statistical Society*, Series B **47**:203-210.
7. Gilmour, A. R., R. D. Anderson and A. L. Rae. 1985. The analysis of binomial data by a generalized linear mixed model. *Biometrika* **72**:593-599.
8. Akaike, H.1973. Information theory and an extension of the maximum likelihood principle. *In Proceedings of the Second International Symposium on Information Theory,* B. *N.* Petrov and F. Csaki (eds), 267-281. Budapest: Akademiai Kiado.
9. 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.
10. 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.
11. 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
12. Gibbons, R.D. and R.D. Bock. 1987. Trend in correlated proportions, Psychometrika **52**:113–124.
13. McCulloch, C.E. 2003. NSF-CBMS Regional Conference series in Probability and Statistics, Volume 7.
14. Rabe-Hesketh, S., A. Skrondal and A. Pickles. 2002. Reliable estimation of generalized linear mixed models using adaptive quadrature. Stata Journal, 2:1-21.
15. Kullback, S. and R.A. Leibler. 1951. On information and sufficiency. Annals of Mathematical Statistics **22**:79-86.
16. Choolani, M. and S.S. Ratnam. 1995. Maternal morbidity: a global overview. Journal of the Indian Medica Association, **93**:36-40.
17. 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.