Journal of Applied Sciences

ISSN 1812-5654

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

Journal of Applied Sciences 5 (7): 1228-1231, 2005
ISSN 1812-5654

Generalized Estimating Equations for Conditional and Unconditional Residuals in Diabetes Mellitus Data

Md. Abdus Salam Akanda, Kawsar Jahan, Maksuda Khanam and M. Ataharul Islam
Department of Statistics, University of Dhaka, Dhaka-1000, Bangladesh

Abstract: This study focused for estimating the parameters of marginal model for repeated binary responses through the Generalized Estimating Equations (GEE) methodology. The GEE were applied to observe how certain covariates relate to change of the disease status overtime. In addition, we focused on the methodology of GEE using conditional and unconditional residuals along with common correlation structures seen in longitudinal studies. Here, the GEE has been applied to the data of four repeated binary observations of the registered patients at BIRDEM. We demonstrate that the estimator of the correlation based on conditional residuals is nearly efficient when compared with maximum likelihood. This estimator also yields more efficient estimates of the correlation than the usual GEE estimator that is based on unconditional residuals. Finally the results of applying the data set are presented.

Key words: Logistic regression, GEE, marginal model, conditional residual, unconditional residual

INTRODUCTION

An increasing popular approach for estimating the parameters of marginal models for repeated binary responses is the GEE methodology. To assess the fit of a model, it is necessary to identify the influential elements. In particular, Liang[1] and Prentice[2] have developed moment-based generalized estimating equations which only require specification of the form of the first two moments of the vector of binary responses for each individual. Instead the modeling the association between the pair of binary responses in terms of the marginal correlations, Lipsitz[3] , Liang[4] and Carey[5] propose using the marginal odds ratio. Carey[5] estimate the marginal odds ratio using conditional residuals and have shown that their estimating equations for the odds ratio are highly efficient when compared to the optimal second-order joint estimating equations (GEE2) of Liang[4]. Carey[5] also demonstrate that there are very significant computational savings from using their method rather than the optimal joint estimating equations. Albert[6] proposed generalized estimating equations for estimating the parameters of both the mean and partial correlation structures. They highlighted on the use of this method for modeling the effect of spatial location and subject-specified covariates on spatially correlated binary data. Albert[7] describe a methodology for jointly modeling the number of events and the vector of correlated binary severity measures. They functionally linked the regression parameters for the counts and binary means and discussed GEE approach for parameter estimation. They also discussed the conditions under which the proposed joint modeling approach provides marked gains in efficiency relative to the common procedure of simply modeling the counts. In this study, we demonstrate that the measures of association between pairs of binary responses, e.g., the parameters can be estimated using conditional residuals and the usual GEE estimator can also be found using unconditional residuals.

Generalized estimating equations: The GEE approach provides consistent estimators of the regression parameters which needs only the correct specification of the form of the mean function of the vector of responses for each individual. In longitudinal studies, there is an implicit ordering of the times of the observations of each individual. We assume that the ith individual is observed at times t = 1,2,.........,T_i, where, T_i need not be the same for all N individuals. With binary response obtained at time t, we form a $T_i \times 1$ vector

$$Y_i = \left[Y_{i1},..........,Y_{iT_i}\right]'$$

where, the binary random variable $Y_{it} = 1$ if the ith individual has response 1 (success) at time t_i and $Y_{it} = 0$ otherwise. Each individual has a J×1 covariate vector x_{it}, measured at time t, which includes both time-stationary and time-varying covariates. Let $X_i = \left[x_{i1},........,x_{iT_i}\right]$

Corresponding Author: Md. Abdus Salam Akanda Lecturer, Department of Statistics, University of Dhaka, Dhaka-1000, Bangladesh Tel: 88029661920-73/4828

represent the $T_i \times J$ matrix of covariates for the ith individual. In the cluster data setting, Y_i is the vector of binary responses for the T_i units within a cluster.
The marginal distribution of Y_{it} is Bernoulli with

$$\pi_{it} = \pi_{it}(\beta) = E\left(Y_{it} \middle| x_{it}, \beta\right) = pr\left(Y_{it}=1 \middle| x_{it}, \beta\right) = \frac{\exp(\theta_{it})}{1+\exp(\theta_{it})} \qquad (1)$$

where, $\theta_{it} = \ln(\pi_{it}/(1-\pi_{it}))$ and β is a $J\times1$ vector of parameters. The $\pi_{it}(\beta)$ can be grouped together to form a vector $\pi_i(\beta)$ containing the marginal probabilities of success, $\pi_{it}(\beta) = E[Y_i/X_i, \beta] = [\pi_{i1}, \ldots\ldots\ldots, \pi i_{Ti}]$. Since Y_{it} is binary, the logistic link function, $\theta_{it} = x'_{it}\beta$, is a natural choice, although, in principle any link function could be chosen.

We are interested in making inference about β, as well as the parameters, say α of the joint distribution of Y_{is} and Y_{it} (Table 1), where:

$$\pi_{ist} = E(Y_{is}Y_{it}|x_{is}, x_{it}, \beta, \alpha) = pr(Y_{is}=1, Y_{it}=1|x_{is}, x_{it}, \beta, \alpha)$$

This joint probability can be modeled in terms of the two marginal probabilities $\pi_{is}(\beta)$ and $\pi_{it}(\beta)$, as well as an association parameter (contained in α). Although the following methods can be used for any association parameter (e.g., marginal odds ratio, kappa coefficient, relative risk), we focus on the marginal correlation coefficient. From Table 1, the correlation between the responses at times s and t is

$$\rho_{ist} = \rho_{ist}(\beta,\alpha) = corr\left(Y_{is}, Y_{it} \middle| x\beta, \alpha\right) = \frac{E\left[\left(Y_{is}-\pi_{is}\right)\left(Y_{it}-\pi_{it}\right)\right]}{\left[\pi_{is}\left(1-\pi_{is}\right)\pi_{it}\left(1-\pi_{it}\right)\right]^{1/2}}$$

In terms of the correlation coefficient, the joint probability π_{ist} can be written as

$$\pi_{ist} = \pi_{is}\pi_{it} + \rho_{ist}\left[\pi_{is}\left(1-\pi_{is}\right)\pi_{it}\left(1-\pi_{it}\right)\right]^{1/2} \qquad (2)$$

In the following, let α denote the parameters of the correlation between pairs of binary responses. Then, to estimate (β,α), we suggest modifying the estimating equations proposed by Carey[5] which were originally developed to estimate the marginal odds ratio. The estimating equations for β are given by:

$$u_\beta\left(\hat{\beta}\right) = \sum_{i=1}^{N} \hat{D}'_i \hat{V}_i^{-1}\left[Y_i - \pi_i\left(\hat{\beta}\right)\right] = 0 \qquad (3)$$

where, $D_i = \delta\pi_i/\delta\beta$ and V_i is the $T_i \times T_i$ "working" covariance matrix of Y_i. The tth diagonal elements of $V_i(\alpha,\beta)$ is $var(Y_{it}) = \pi_{it}(1-\pi_{it})$, which is specified entirely by the marginal distributions i.e., by β. The st th off-diagonal elements of V_i is $cov(Y_{is}, T_{it}) = \pi_{its} - \pi_{is}\pi_{it}$, where π_{its} is specified by Eq. 2.

Table 1: Cross-classification probabilities for times s and t, s≠t

		Time t		
		1	2	
Time (sec)	1	π_{ist}	$\pi_{is}-\pi_{ist}$	π_{is}
	0	$\pi_{it}-\pi_{ist}$	$1-\pi_{is}-\pi_{it}+\pi_{ist}$	$1-\pi_{is}$
		π_{it}	$1-\pi_{it}$	1.0

If α is unknown (which is typically the case), then it must be estimated with a set of estimating equations similar to (3). Following Carey[5] for a pair of times s<t we form the conditional residuals $\{Y_{it} - E(Y_{it})|Y_{is} = y_{is}, X_i\}$, that is, deviations about conditional expectations. These random variables can than be grouped together to form the $[T_i(T_i-1)/2]\times1$ vector of conditional residuals, $(U_i - \eta_i)$, where:
$U_i = \{U_{i12}, U_{i13}, \ldots\ldots, U_{i\,(Ti-1)\,Ti}\}'$, $\eta_i = \{\eta_{i12}, \eta_{i13}, \ldots\ldots, \eta_{i(Ti-1)Ti}\}'$, with $U_{ist} = Y_{it}$ and $\eta_{ist} = E(Y_{it}|Y_{is} = y_{is}, X_i)$, for s<t.
From Table 1,

$$\eta_{ist}(\beta,\alpha) = E\left(Y_{it} \middle| Y_{is}=y_{is}, \beta, \alpha\right) = y_{is}\left[\frac{\pi_{ist}}{\pi_{is}}\right] + \left(1-y_{is}\right)\left[\frac{\pi_{it}-\pi_{ist}}{1-\pi_{is}}\right] \qquad (4)$$

In order to form another set of moment estimating equations similar to (3), we need to take appropriate linear combinations of $[U_i - \eta_i]$. Thus a second set of (moment) estimating equations for α is given by:

$$u_a\left(\hat{\alpha}\right) = \sum_{i=1}^{N} \hat{C}'_i \hat{W}_i^{-1}\left[U_i - \eta_i\left(\hat{\beta}, \hat{\alpha}\right)\right] = 0 \qquad (5)$$

where, $C_i = \delta\eta_i/\delta\alpha$ and $W_i = diag\{var(Y_{it}|Y_{is} = y_{is})\}$ with $var(Y_{it}|Y_{is} = y_{is}) = \eta_{ist}(1-\eta_{ist})$. Using these estimating equations, the estimate $\left(\hat{\beta}, \hat{\alpha}\right)$ is the solution to (3) and (5) and can be obtained using a Gauss-Seidel algorithm.
Using Taylor series expansions similar to Prentice[2] assuming that regression for Y_i and the model for the association has been correctly specified, $\left(\hat{\beta}, \hat{\alpha}\right)$ is consistent for (β,α). In addition, $N^{1/2}\left\{\left(\hat{\beta}-\beta\right), \left(\hat{\alpha}-\alpha\right)\right\}$ has an asymptotic distribution which is multivariate normal with mean vector 0. In contrast to (5), Prentice[2] forms the unconditional residuals $\{Y_{is}Y_{it} - E(Y_{is}Y_{it}|X_i)\}$, that is, the deviations about unconditional expectations. These random variables can then be grouped together to form $[T_i(T_i-1)/2]\times1$ vector, $(p_i - v_i)$, where
$P_i = \{P_{i12}, P_{i13}, \ldots\ldots\ldots, P_{i\,(Ti-1)\,Ti}\}'$, $v_i = \{v_{i12}, v_{i13}, \ldots\ldots\ldots, v_{i(Ti-1)Ti}\}'$, with $P_{ist} = (Y_{is}, Y_{it})$ and $v_{ist} = E(Y_{is}, Y_{it}|X_i)$, for s<t. Then, Prentice[2] proposes the following set of (moment) estimating equations for α,

$$u_a(\hat{\alpha})=\sum_{i=1}^{N}\hat{A}_i'\hat{H}_i^{-1}\left[P_i - v_i\left(\hat{\beta},\hat{\alpha}\right)\right]=0 \qquad (6)$$

where, $A_i = \delta v_i / \delta\alpha$ and $P_i \approx \mathrm{cov}(P_i)$. Prentice[2] also suggests specifying the "working" covariance matrix for P_i as $\mathrm{diag}\{\mathrm{var}(P_{ist})\}$, where $\mathrm{var}(P\ )_{ist}= \pi_{ist}(1-\pi_{ist})$ since (Y_{is}, Y_{it}) is binary. Assuming that the regression for Y_i and the model for the association has been correctly specified, the estimating equations proposed by Prentice[2] yield estimate $\left(\hat{\beta},\hat{\alpha}\right)$ that are consistent for (β,α). In addition, $N^{1/2}\left\{\left(\hat{\beta}-\beta\right),\left(\hat{\alpha}-\alpha\right)\right\}$ has an asymptotic distribution which is multivariate normal with mean vector zero.

Data and variables: In our study we have used the repeated measures data diabetes mellitus to carry out the analysis. Here the follow up data on 528 patients registered at BIRDEM (Bangladesh Institute of Research and Rehabilitation in Diabetes, Endocrine and Metabolic disorders) in 1984-94 are used to identify the risk factors responsible for the transitions from controlled diabetic to confirmed diabetic state as well as confirm diabetic to controlled stage of diabetes. We have taken into account the four consecutive visits of the patients from the registration. The response variable is defined in terms of the observed glucose level 2 h of 75 g glucose load for each follow-up visit. The cut-off point for the blood glucose level is 11.1 m mol L^{-1}. If the observed response is less than 11.1, then the patient is defined as non diabetic (categorized as 0) if the response is greater than or equal to 11.1 then the patient is said to be diabetic (categorized as 1) according to WHO (1985) criteria. We include six independent variables in this study. They are age, sex, education level, area of residence, family history of father and mother and time. Out of these variables, age represents the age of the respondents at each visit. Time represents the length of time of the consecutive visits. These two variables are continuous variables and used directly in the analysis. Sex, education level, area of residence and family history of father and mother are categorical variables. Here sex is a dichotomous variable with two categories 0 and 1, 0 stands for female and 1 stands for male.

Education level is categorized again 0 and 1. Here, 0 represents the patients having below secondary education and 1 represents the patients having the secondary education or more. Area has two categories, 0 represents rural and 1 represents urban or semi-urban. FHFM represents the genetic history of the parents. This variable has two categories, 0 representing the non-diabetic father and mother and 1 representing anyone or father and mother diabetic.

RESULTS AND DISCUSSION

The logistic regression model is considered as one of the most important and widely applicable techniques in analyzing repeated outcome variables. To assess the fit of a model, it is necessary to identify the influential elements. In the logistic regression analysis for repeated binary measures we adjust for setting and the covariates. We assumed independence, exchangeable and autoregressive working correlation structures and we obtained standard errors.

These analyses were carried out using specially written S-plus program and results shown in Table 2 and 3. We found that for the repeated binary responses, the variables education level, area, family history of father and mother (i.e., the disease status of the parents) and time are significant under independence, exchangeable and autoregressive correlation assumptions and thus have considerable effect in changing the disease status. We also found that under all assumptions education level and area shows negative association and Family History of Father and Mother (FHFM), time shows positive association. Among these variable education level, area and time are significant at 5% level of significance in all cases (GEE for conditional and unconditional residuals) (Table 2 and 3). The only variable Family History of Father and Mother (FHFM) is significant at 10% level of significance in all cases. The estimated coefficients of the variables age and sex are found to be insignificant in all cases. Hence it may be conclude that the variables age and sex has no significant effect on the transition from confirmed diabetes state to controlled diabetes state.

Table 2: Estimates obtained by GEE assuming the various correlation structures within repeated outcomes with associated Wald test statistic for conditional residuals

	MLE			Exchangeable			Autoregressive		
Parameters	Estimates	Wald statistic	Odds ratio	Estimates	Wald statistic	Odds ratio	Estimates	Wald statistic	Odds ratio
Constant	0.2959	1.0155	1.3443	0.3093	0.9891	1.3625	0.3396	1.0466	1.4044
Age	0.0005	0.1134	1.0005	0.0005	0.0942	1.0005	0.0003	0.0669	1.0003
Sex	-0.0872	-0.7822	0.9165	-0.0860	-0.7189	0.9176	-0.0756	-0.6101	0.9272
Edlv	-0.3679	-2.9976	0.6922	-0.3659	-2.7769	0.6936	-0.3579	-2.6161	0.6991
Area	-0.3263	-2.7031	0.7216	-0.3289	-2.5381	1.7197	-0.3282	-2.4369	0.7202
FHFM	0.2032	1.7144	1.2253	0.2071	1.7272	1.2301	0.2006	1.7143	1.2221
Time	0.0819	3.5754	1.0855	0.0705	3.0385	1.0729	0.0795	3.2906	1.0828
	Likelihood ratio=172.54			Likelihood ratio=179.166			Likelihood ratio=193.368		

Table 3: Estimates obtained by GEE assuming the various correlation structures within repeated outcomes with associated Wald test statistic for unconditional residuals

	MLE			Exchangeable			Autoregressive		
Parameters	Estimates	Wald statistic	Odds ratio	Wald Estimates	Wald statistic	Odds ratio	Estimates	Wald Estimates	Odds ratio
Constant	0.2959	1.0155	1.3443	0.3123	1.0368	1.3666	0.3453	1.0947	1.4124
Age	0.0005	0.1134	1.0005	0.0004	0.0894	1.0004	0.0003	0.0672	1.0003
Sex	-0.0872	-0.7822	0.9165	-0.0796	-0.6895	0.9235	-0.0733	-0.6122	0.9294
Edlv	-0.3679	-2.9976	0.6922	-0.3612	-2.8395	0.6968	-0.3457	-2.6132	0.7077
Area	-0.3263	-2.7031	0.7216	-0.3285	-2.6267	0.7200	-0.3241	-2.4909	0.7232
FHFM	0.2032	1.7144	1.2253	0.2044	1.6636	1.2267	0.2041	1.6868	1.2265
Time	0.0819	3.5754	1.0855	0.0721	3.2228	1.0748	0.0794	3.4012	1.0826
	Likelihood ratio=172.54		Likelihood ratio=180.401		Likelihood ratio=193.84				

Table 4: Asymptotic relative efficiency of the GEE Estimator based on conditional and unconditional residuals relative to the MLE

	Unconditional residual		Conditional residual	
Variables	R.E.(Exchangeable)	R.E.(Autoregressive)	R.E.(Exchangeable)	R.E.(Autoregressive)
Age	1.038	1.077	1.0750	1.115
Sex	1.036	1.073	1.0730	1.111
Edlv	1.037	1.076	1.0740	1.115
Area	1.036	1.077	1.0730	1.116
FHFM	1.037	1.072	1.3074	1.111
Time	0.976	1.018	1.0110	1.054

From Table 4, we found the asymptotic efficiency of the GEE estimator assuming exchangeable and autoregressive correlation relative to the ML method. Comparing the results we come to the conclusion that parameters are estimated more efficiency by the GEE estimator based on conditional residuals than the unconditional residuals. Under the assumption of autoregressive correlation structure the asymptotic relative efficiency is more than other correlation structures.

CONCLUSIONS

From the present data set it can be seen that parameter estimates based on both conditional and unconditional residuals are more efficient than the ML estimates. We may conclude that for analyzing the data in case of chronic disease (i.e., diabetic mellitus), where the response variable is binary and the resulting estimates of GEE based on conditional residuals can be used more efficiently under the assumption of autoregressive correlation than that based on unconditional residuals. Furthermore, estimating equations based on conditional residuals could be constructed to estimate the association of repeated ordinal data. We conjecture that these estimating equations will also be more efficient than the estimating equations based on unconditional residuals in the present data set.

ACKNOWLEDGMENT

We would like to express my gratitude to the Director of BIRDEM for giving us kind permission to use their data. We are indebted to the Chairman, Department of Statistics, University of Dhaka, Bangladesh for his kind cooperation through this research.

REFERENCES

1. Liang, K.Y. and S.L. Zeger, 1986. Longitudinal data analysis using generalized linear models. Biometrics, 73: 13-22.
2. Prentice, R.L., 1988. Correlated binary regression with covariates specific to each binary observation. Biometrics, 44: 1033-1048.
3. Lipsitz, S.R., N.M. Laird and D.P. Harrington, 1991. Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association. Biometrica, 78: 153-160.
4. Liang, K.Y., S.L. Zeger and B. Qaqish, 1992. Multivariate regression analysis for categorical data (with discussion). J. Royal Statistical Society, Series B, 54: 3-40.
5. Carey, V., S.L. Zeger and P.J. Diggle, 1993. Modeling multivariate binary data with alternating logistic regressions. Biometrica, 80: 517-526.
6. Albert, P.S. and L.M. Shane, 1995. A generalized estimating equation approach for spatially correlated binary data: Applications to the analysis of neuroimaging data. Biometrics, 51: 627-638.
7. Albert, P.S., D.A. Follmann and H.X. Barnhart, 1997. A generalized estimating equation approach for modeling random length binary vector data. Biometrics, 53: 1116-1124.