Journal of Applied Sciences 5 (7): 1214-1218, 2005
ISSN 1812-5654

Comparison of Goodness-of-Fit Tests for GEE Modeling with Binary Responses to Diabetes Mellitus

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

Abstract: Analysis of data with repeated measures is often accomplished through the use of Generalized Estimating Equations (GEE) methodology. Although methods exist for assessing the adequacy of the fitted models for uncorrelated data with likelihood methods, it is not appropriate to use these methods for models fitted with GEE methodology. Barnhart and Williamson[1] proposed model-based and robust (empirically corrected) goodness-of-fit tests for GEE modeling with binary responses based on partitioning the space of covariates into distinct regions and forming score statistics that are asymptotically distributed as chi-square random variables with the appropriate degrees of freedom. In their suggested GEE approach the correlation between two responses was not considered. We here proposed an alternative procedure based on GEE where the correlation between two responses was considered. We extended their work using different correlation structures exchangeable, autoregressive and pairwise correlation along with their suggested identity correlation structure.

Key words: Correlated data, GEE modeling, goodness-of-fit test, logistic regression, score test

INTRODUCTION

The use of Generalized Estimating Equations (GEE) to analyze repeated binary data has become increasingly common in the health sciences. The analysis of correlated binary responses is often accomplished through the use of GEE methodology for parameter estimation. Assessment of the adequacy of the fitted GEE model is problematic since no likelihood exists and the residuals are correlated within a cluster. Tsiatis[2] proposed a goodness-of-fit test for the logistic regression model which is asymptotically chi-squared and is computed as a quadratic form of observed counts minus the expected counts. Stuart[3] proposed a goodness-of-fit test statistic for regression with heterogeneous variance, which is asymptotically chi-square if the given model is correct. The test statistic is computed as a quadratic form of observed minus predicted responses. Cessie[4] discussed a new global test statistic for models with continuous covariates and binary response is introduced. The test statistic is based on nonparametric kernel methods. Explicit expressions are given the mean and variance of the test statistic. Asymptotic properties are considered and approximate corrections due to parameter estimation are presented. Also Cessie[5] considered testing the goodness-of-fit of regression models. Emphasis is on a goodness-of-fit test for generalized linear models with canonical link function and known dispersion parameter. The test based on the score test for extra variation in a random effect model. By choosing a suitable form for the dispersion matrix, a goodness-of-fit test statistic is obtained which is quite similar to test statistics based on non-parametric kernel methods. The aim of present study was to utilize the BIRDEM data to parameter estimate in the main effect model and another model which includes the same main effects, the regions, time effects and interaction effects and then to test the goodness-of-fit by using various correlation structures.

Generalized Estimating Equation (GEE): The GEE approach provides consistent estimators of the regression parameters which needs only the correct specification of the form of the mean function μ_i, of the vector of responses for each individual.

Let us consider that each individual is observed for T occasions. Thus we have a Y x 1 random vector of responses for the ith individual where the response variable is binary. Notationally,

$$Y_i = \begin{bmatrix} Y_{i1} & Y_{i2} & \cdots & Y_{iT} \end{bmatrix}'$$

Where, the binary random variable $Y_{it} = 1$ if at time t, the subject i has response 1, i.e., success and 0 otherwise. Here the response variable is dichotomous. We took k independent variables, so for ith individual we have a T x k matrix of covariates.

Corresponding Author: Md. Abdus Salam Akanda, Department of Statistics, University of Dhaka, Dhaka-1000, Bangladesh
Tel: (88 02) 9661920-73/4828 Fax: (88 02) 865583 (Office)

Notationally,

$$X_i = \begin{bmatrix} X_{i11} & X_{i12} & \cdots & \cdots & X_{i1k} \\ X_{i21} & X_{i22} & \cdots & \cdots & X_{i2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{iT1} & X_{iT2} & \cdots & \cdots & X_{iTk} \end{bmatrix}_{T\times k}$$

$$= \begin{pmatrix} X_{i1} & X_{i2} & \cdots & X_{iT} \end{pmatrix}^T$$

Where, $X_{ij} = \begin{pmatrix} X_{ij1} & X_{ij2} & \cdots & X_{ijk} \end{pmatrix}; j = 1, 2, \ldots\ldots, T$

The usual GEE modeling for binary outcomes have the following setting:

$$\text{log it}\left(\mu_i\right) = X_i\beta \tag{1}$$

The mean vector is

$$\mu_i = \begin{bmatrix} \mu_{i1} \\ \mu_{i2} \\ \vdots \\ \mu_{iT} \end{bmatrix}_{T\times 1} = \begin{bmatrix} E(Y_{i1}) \\ E(Y_{i2}) \\ \vdots \\ E(Y_{iT}) \end{bmatrix} = \begin{bmatrix} p_{i1} \\ p_{i2} \\ \vdots \\ p_{iT} \end{bmatrix} = P_i;$$

Where:

$$\mu_{ij} = p_{ij} = \Pr(y_{ij} = 1/X_{ij}),\ j = 1, 2, \ldots\ldots, T;\ i = 1, 2, \ldots\ldots, N.$$

$$1-p_{ij}=1-\mu_{ij}.$$

So the variance of y_{ij} is $p_{ij}\left(1-p_{ij}\right) = \mu_{ij}\left(1-\mu_{ij}\right)$

And the variance covariance matrix of y_i is given by:

$$V(Y_i) = \begin{bmatrix} V(Y_{i1}) & Cov(Y_{i1},Y_{i2}) & \cdots & Cov(Y_{i1},Y_{iT}) \\ Cov(Y_{i1},Y_{i2}) & V(Y_{i2}) & \cdots & Cov(Y_{i2},Y_{iT}) \\ \vdots & \vdots & \vdots & \vdots \\ Cov(Y_{i1},Y_{iT}) & Cov(Y_{i2},Y_{iT}) & \cdots & V(Y_{iT}) \end{bmatrix}_{T\times T}$$

Estimation of β is obtained by solving the generalized estimating equations[6,7],

$$\sum_{i=1}^{N}\left(\frac{\delta\mu_i}{\delta\beta_p}\right)' V_i^{-1}\left(Y_i-\mu_i\right) = 0, \quad p = 1, 2, \ldots\ldots, P+1 \tag{2}$$

with $V_i = A_i^{1/2} R_i A_i^{1/2}$, $A_i = \text{diag}\left(\text{var}\left(y_{i1}\right), \ldots\ldots, \text{var}\left(y_{iT}\right)\right)$, where, R_i is the working correlation matrix for Y_i.

Goodness-of-fit test: By first partitioning the covariate space $X = (x_1, x_2, \ldots\ldots, x_p)'$ into M distinct region in P-dimensional space. Let be an $I_{it} = (I_{it1}, I_{it2}, \ldots\ldots, I_{itM})'$ be an M x 1 vector, where, I_{itm} is the indicator variable that equals one if the ith subject is in the mth region at the tth occasion and zero otherwise. They define the T x M matrix I_i as:

$$I_i = \left[I_{i1}, I_{i2}, \ldots\ldots, I_{iT}\right]' \tag{3}$$

Let Z_T be the T x (T-1) matrix where the first row has entries zero and the remaining (T-1) rows form a (T-1) x (T-1) identity matrix. Consider the model :

$$\text{log it}\left(\mu_i\right) = X_i\beta+Z_T\tau+I_i\gamma+S_i\rho \tag{4}$$

Where, $S_i = [0, \text{diag}\,(I_{i2}, I_{i3}, \ldots\ldots, I_{iT})']$ is a T x (T-1) M matrix and 0 is a (T-1) M x 1 vector of zeros. Note that τ is the (T-1) x 1 vector of time effects (the first occasion is the reference time point), γ is the M x 1 vector of region effects and ρ is the (T-1) M x 1 vector of time and region interaction effects because each column of S_i results from component wise multiplication of two column vectors, one column vector from Z_T and the other from I_i. A goodness-of-fit statistic consists of testing H_0: $\theta = 0$, where, $\theta = [\tau', \gamma', \rho']'$ is a J x 1 vector with J = (T-1)+M+(T-1)M.

Let L = P+1+J be the number of parameters in the model presented in (4). Denote U be the L x 1 vector with lth component:

$$U_l = \sum_{i=1}^{N}\hat{D}_{il}'\hat{V}_i^{-1}(Y_i-\hat{\mu}_i) \tag{5}$$

for $l = 1, 2, \ldots\ldots\ldots, L$, where, $\hat{D}_{il} = \partial\hat{\mu}_i/\partial\beta_l$ for $l \le p+1$,

$\hat{D}_{il} = \delta\hat{\mu}_i/\delta\theta_{l-p-1}$ for $l > P+1$,

$\hat{\mu}_i = \text{log it}^{-1}\left(X_i\hat{\beta}+Z_T\tau+I_i\gamma+S_i\rho\right)$ and $\hat{\beta}$

is obtained as the solution to (2). Then under H_0: $\theta = 0$, the asymptotic distribution of U is multivariate normal with mean zero and covariance matrix[6]:

$$W_R = \sum_{i=1}^{N}\hat{D}_i'\hat{V}_i^{-1}\text{Cov}(Y_i)\hat{V}_i^{-1}\hat{D}_i \tag{6}$$

Where, $\hat{D}_i = [\hat{D}_{il}, \ldots\ldots, \hat{D}_{iL}]$ is a T x T matrix. Note that cov (Y_i) can be consistently estimated by $(Y_i-\hat{\mu}_i)(Y_i-\hat{\mu}_i)'$[6]. If the correlation matrix R_i is correctly specified, then the asymptotic covariance matrix U reduces to $W = \sum_{i=1}^{N}\hat{D}_i'\hat{V}_i^{-1}\hat{D}_i$

Let $U = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}$ $W_R = \begin{pmatrix} A_R & B_R' \\ B_R & C_R \end{pmatrix}$ $W = \begin{pmatrix} A & B' \\ B & C \end{pmatrix}$

be the partitioning for U, W_R and W, where, U_2 is the J x 1 vector and C_R and C are J x J matrices. Under H_0: θ=0, both the proposed robust (empirically corrected) goodness-of-fit test statistic:

$$Q_R = U_2'(C_R - B_R A_R^{-1} B_R')^{-} U_2$$

And the proposed model-based goodness-of-fit test statistic:

$$Q = U_2'(C - B A^{-1} B')^{-} U_2$$

Are asymptotically distributed as chi-square random variables with:

$$d.f = \text{rank}((C_R - B_R A_R^{-1} B_R')^{-}) = \text{rank}((C - BA^{-1}B')^{-}),$$

Where, G^{-} is any generalized inverse of the matrix G. The degree-of-freedom for chi-square random variables do not equal the number of parameters in θ because of linear dependencies between the covariates in the model and the covariates from the region partitioning, i.e., $(C_R - B_R A_R^{-1} B_R')$ and $(C - BA^{-1}B')$ are singular matrices. Let H_1 and H_2 be the design matrices in models (1) and (4), respectively.

Then intuitively, the degrees-of-freedom of the above chi-square random variables is equal to rank (H_2)-(H_1). Let $H_{2i} = \{h_{itj}\} = [X_i, Z_T, I_i, S_i]$ be the $T \times (P+1+J)$ design matrix for the ith subject in model (4). It is easily shown that the tj th element of $\hat{D}_i$ is equal to $\hat{\mu}_{it}(1-\hat{\mu}_{it})h_{itj}$. Therefore, the goodness-of-fit test statistics Q and Q_R can be readily calculated once $\hat{\beta}$ is obtained from the estimating Eq. 2.

Data set and covariates: In our study we have used the repeated measures data diabetes mellitus to carry out the analysis. Here the follow up data on 995 patients registered at BIRDEM (Bangladesh Institute of Research and Rehabilitation in Diabetes, Endocrine and Metabolic disorders) in 1984-94 is 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. The response variable is defined in terms of the observed glucose level two hours of 75 g-glucose load follow-up visit. The cut-off point for the blood glucose level is 11.1 mmol L^{-1}. If the observed response is less than 11.1, then the patient is define 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). We included two independent variables in the study. They are age and sex. Out of these variables, age represents the age responds at each visit. The variable is a continuous variable and used directly in the analysis. Sex is categorical variables. Here sex is a dichotomous variable with two categories 0 and 1, 0 stands for female and 1 stands for male. In order to assess the performance of the proposed goodness-of-fit tests, we used data simulated with known distributions from models in the alternative hypothesis to test the goodness-of-fit. To conduct the proposed goodness-of-fit tests, the following regions were partitioned as region1 if age greater than or equal to 50 and male, region 2 if age greater than or equal to 50 and female, region 3 if age less than 50 and male and region 4 if age less than 50 and female. If any individual occurs any of the four regions then indicate 1 otherwise 0. Time effect represents the two consecutive visits. Time effect is a dichotomous variable with two categories 0 and 1, 0 stands for first visit and 1 stands for second visit. Interaction 1, interaction 2, interaction 3, interaction 4 are component wise multiplication of region 1, region 2, region 3, region 4 and time effect.

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, autoregressive and pairwise working correlation structures and we obtained standard errors. Table 1 lists the parameter estimates and standard errors for the initial model having only main effects.

According to likelihood test the null hypothesis is rejected under all correlation structures in GEE. In this case has an interpretation that at least one of the coefficients is different from zero. According to Wald test sex is significant at 5% level of significance under independence, exchangeable, autoregressive and pairwise correlation structures. There exits positive association between the response variable and sex. The estimated coefficient of the variable age is found to be insignificant in all cases. Hence it may be conclude that these variables has no significant effect on the transition from confirmed diabetes state to controlled diabetes state. In terms of odds ratio, we may comment that, male patients are 1.240775 times likely to develop diabetes as compared to their counterparts. We considered additions to this main effects model to provide a better fit to the data. Table 2 displays the results from a model that includes regions, time effects and interactions.

In this case we see that several of the effects are significant, indicating their importance in modeling. Reject the null hypothesis by likelihood test under independence, exchangeable autoregressive and pairwise correlation structures. So rejection of null hypotheses in this case has an interpretation that at least one of the coefficients is different from zero. We also found that under all assumptions region 1 and time effect show

Table 1: Estimates obtained by GEE assuming various correlation structures within repeated outcomes with associated Wald test

	Independence		Exchangeable		Autoregressive		Pairwise	
Parameter	Estimate	Wald statistic	Estimate	Wald statistic	Estimate	Wald statistic	Estimate	Wald statistic
Intercept	-0.05133	-0.23399	-0.05301	-0.23151	-0.06136	-0.26059	-0.06429	-0.25944
Age	-0.00429	-1.07549	-0.00467	-1.09089	-0.00448	-1.08699	-0.00499	1.11437
Sex	0.21574	2.57915*	0.24637	2.75163*	0.23101	2.40203*	0.25094	2.37358*
	Likelihood ratio=71.99052		Likelihood ratio=72.6174		Likelihood ratio=73.4638		Likelihood ratio=85.395	

*Significant at $p<0.05$

Table 2: Estimates obtained Barnhart and Williamson's model by GEE assuming various correlation structures within repeated outcomes with associated Wald test

	Independence		Exchangeable		Autoregressive		Pairwise	
Parameter	Estimate	Wald statistic	Estimate	Wald statistic	Estimate	Wald statistic	Estimate	Wald statistic
Region 1	0.31925	2.9088*	0.31925	2.9087*	0.33101	2.4309*	0.34263	2.3565*
Region 2	-0.00239	-0.6853	-0.00239	-0.6853	-0.00236	-0.6271	-0.00249	-0.6387
Region 3	-0.00449	-1.0464	-0.00449	-1.0464	-0.00460	-0.9850	-0.00473	-0.9875
Region 4	-0.00449	-1.0973	-0.00449	-1.0973	-0.00466	-1.0909	-0.00478	-1.0887
Time effect	0.21989	2.3146*	0.21989	2.3146*	0.24595	2.3105*	0.23863	2.0659*
Interaction 1	-0.53979	-2.1587*	-0.53979	-2.1587*	-0.55632	-2.0853*	-0.56792	-2.1148*
Interaction 2	-0.00090	-0.5304	-0.00089	-0.5304	-0.00094	-0.5423	-0.00104	-0.5973
Interaction 3	0.00699	0.1749	0.00699	0.1749	0.00758	0.1626	0.00814	0.1588
Interaction 4	-0.00599	-0.1927	-0.00599	-0.1927	-0.00629	-0.1647	-0.00716	-0.1711
	Likelihood ratio=264.58		Likelihood ratio=268.81		Likelihood ratio=271.14		Likelihood ratio=275.57	

Table 3: Goodness-of-fit by using various correlation structures

	Independence		Exchangeable		Autoregressive		Pairwise	
Test	Statistic	p-value	Statistic	p-value	Statistic	p-value	Statistic	p-value
Model based (Q)	28.56	0.00077	27.87	0.00100	27.42	0.00119	25.53	0.00244
Empirically corrected (Q_R)	25.34	0.00262	24.69	0.00333	24.31	0.00384	21.86	0.00933

positive association and interaction1 shows negative association. Among these variation region1, time effect and interaction1 are significant at 5% level of significance in all cases. The other coefficients of the variables are found to be insignificant in all cases. Hence it may be conclude that these variables has no significant effect on the transition from confirmed diabetes state to controlled diabetes state.

From the Table 3, the model suggested by Barnhart and Williamson[1] is highly significant by model based test. In this case has an interpretation that at least one of the coefficients is different from zero. Also we see that the null hypothesis is rejected by the empirically corrected test and the model (4) is highly significant. In this case has an interpretation that the covariates have significant effect. The both goodness-of-fit test provided no evidence for lack of fit by adding regions, time effect and interaction effects.

CONCLUSIONS

We fit two models to the data. The first model only includes the main effects of age and sex and the second model includes the same main effects and the treatment and time interaction. Because all the covariates are discrete, the covariate categories were used to form four regions with frequencies. Both the goodness-of-fit tests suggest that the model with only main effects did not fit the data well. There is a significant time and treatment interaction effect indicating that patients with new treatment improved significantly faster than the patients with the standard treatment. The model with this interaction term included has a good fit to the data. The parameter estimates and the goodness-of-fit tests obtained here are very similar to the results obtained by using a weighted least squares approach. Thus, the goodness-of-fit tests successfully detected the interpretation departure and the efficiencies of the estimates of the Barnhart and Williamson's suggested model for identity correlation is higher than that of our suggested exchangeable correlation, autoregressive correlation and pairwise correlation.

ACKNOWLEDGEMENTS

We would like to express our 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. Barnhart, H.X. and J.M. Williamson, 1998. Goodness-of-fit test for GEE modeling with binary responses. Biometrics, 54: 720-729.
2. Tsiatis, A.A., 1980. A note on a goodness-of-fit test for the logistic regression model. Biometrika, 67: 250-251.
3. Lipsitz, S.R. and J.F. Buoncristiani, 1994. A robust goodness-of-fit test statistic with application to ordinal regression models. Statistics in Medicine, 13: 143-152.
4. Cessie, S.L. and J.C. Van Houwelingen, 1991. A goodness-of-fit test for binary regression models based on smoothing models. Biometrics, 47: 1267-1282.
5. Cessie, S.L. and J.C. Van Houwelingen, 1995. Testing the fit of a regression model via score tests in random effects models. Biometrics, 51: 600-614.
6. Liang, K.Y. and S.L. Zeger, 1986. Longitudinal data analysis using generalized linear models. Biometrics, 73: 13-22.
7. Zeger, S.L. and K.Y. Liang, 1986. Longitudinal data analysis for discrete and continuous outcomes. Biometrics, 42: 121-130.