Journal of Applied Sciences

ISSN 1812-5654

science alert

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

Journal of Applied Sciences 5 (9): 1597-1601, 2005
ISSN 1812-5654

Goodness-of-Fit Tests for GEE Models Using Kappa-like Statistic to Diabetes Mellitus Study

[1]Md. Abdus Salam Akanda and [2]Maksuda Khanam
[1]Department of Statistics, University of Dhaka, Dhaka-1000, Bangladesh
[2]Mirza Abbas Mohila Degree College, Shahjahanpur, Dhaka-1217, Bangladesh

Abstract: A kappa-like classification statistic is used for assessing the fit of GEE regression models with a categorical response. The statistic is a summary measure depicting how well categorical responses are predicted from the fitted GEE model. The statistic takes on a value of 1 if prediction is perfect and a value of 0 if the fitted model fares no better than random chance, i.e., fitting the repeated categorical responses with an intercept-only model. To assess the performance of the classification statistic, we conducted analyses by using BIRDEM data and the concern is assessing the fit of the GEE categorical response models by determining how well the covariates predict the subject's responses.

Key words: Cumulative logistic regression, kappa statistic, logistic regression, ordinal response

INTRODUCTION

Generalized Estimating Equations (GEE) are useful for analyzing correlated data with categorical or continuous responses [1,2]. Parameter estimation is conducted through estimating equations which converge to a sum of mean zero random variables if the mean structure is correctly specified. There is no need to specify a joint distribution for the responses. However, assessing model fit is further complicated with GEE than for models assuming independence because no likelihood is available and the residuals are correlated within a cluster. Some methods are available for assessing the fit of GEE regression models with binary responses. Horton[3] developed a goodness-of-fit test for assessing such model fit by extending Hosmer[4] goodness-of-fit statistic for ordinary logistic regression. Their proposed statistic has an approximate chi-squared distribution when the model is specified correctly. Barnhart[5] also propose a goodness-of-fit statistic for assessing the fit of GEE binary regression models. They extend Tsiatis' method[6] for assessing the fit of ordinary logistic regression models. This approach involves 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. Barnhart's[5] approach is best employed in the situation when there are only discrete covariates available because then there is no need to partition the covariates. Pan[7] has proposed goodness-of-fit tests for GEE with correlated binary data. Pan's two tests result in the Pearson chi-square and an unweighted sum of residual squares, both of which are based on the residuals. These two tests can only be used when there is at least one continuous covariate available. If the possibility of influential observations is of concern to the data analyst, Preisser[8] have proposed deletion diagnostics for generalized estimating equations. The diagnostics consider leverage and residuals to measure the influence of a subset of observations on the fitted regression parameters. Preisser[9] also generalize the GEE procedure to produce parameter estimates and fitted values that are resistant to influential data. Here, the concern is assessing the fit of GEE categorical response models by determining how well the covariates predict the subject's responses. We present the kappa-like classification statistic, which indicates how well the proposed model predicts the categorical response.

GENERALIZED ESTIMATING EQUATION (GEE)

We outline Lipsitz[10] method as follows. Assume the response of interest is a categorical outcome with K categories denoted $z_{it} = k$ if the tth subunit from the ith cluster falls in the kth category, for i = 1, 2,.....,N; t = 1, 2,.....,T_i($T = \max(T_i) \forall i = 1, 2,.....,N$) and k = 1, 2,.....,K. For simplicity, we will assume that the data are balanced, i.e., $T_i = T$ for i = 1, 2,.....,N. The following method will still be applicable in the case of unbalanced data. The T(K-1)x1

Corresponding Author: Md. Abdus Salam Akanda, Department of Statistics, University of Dhaka, Dhaka-1000, Bangladesh
Fax: (88 02) 8615583

response vector Y_t for cluster I consists of the binary random variables Y_{itk}, where, $Y_{itk} = 1$ if $Z_{it} = k$.

Typically one models the marginal cumulative probabilities of response, $v_{itk} = Pr(Z_{it} \leq k)$ for k = 1, 2,.....,K-1. The marginal probabilities are denoted by $\pi_{itk} = Pr(Z_{it} = k) = Pr(Y_{itk} = 1) = E(Y_{itk}) = v_{itk} - v_{it,k-1}$ and will comprise the T(K-1)x1 vector π_i. The vectors Y_i and π_i require only T(K-1) elements instead of TK elements because $\sum_{k=1}^{K} Y_{itk} = \sum_{k=1}^{K} \pi_{itk} = 1$, for i = 1, 2,.....,K-1 and t = 1,2,.....,T. Let X_{it} be the p x 1 covariate vector for the tth subunit of the ith cluster.

The cumulative marginal response probabilities will be related to the covariates via ling function g, the kth intercept λ_k and the px 1 marginal parameter vector β, $g(v_{itk}) = \lambda_k + X'_{it}\beta$.

The intercept are in increasing order: $\lambda_1 \leq \lambda_2 \leq ..... \leq \lambda_{K-1}$. For an ordinal response, the function g may be any link function such as the logit function, probit function (ϕ^{-1}), or the elementary log-log function. Lipsitz[10] suggest that one estimates β with the following set of generalized estimating equations:

$$v_1(\beta) = \sum_{i=1}^{N} D_i' V_i^{-1} (Y_i - \pi_i) = 0,$$

where, $D_i = D_i(\beta) = d\pi_t(\beta)|d\beta$, $Vi = Vi(\beta, \alpha) \approx var(Y_i)$ is a working covariance matrix of Y_i[1,2] and α is a q x 1 vector of correlation parameters. The parameters α are associated with the correlation between the elements of the vectors Y_{is} and Y_{it}.

KAPPA-LIKE STATISTIC

We used a kappa-like statistic to assess model fit for GEE categorical response models. Historically, the kappa coefficient has been used to determine the agreement of binary[11] and categorical[12] outcomes between raters. Kappa corrects the percentage of agreement between raters by taking into account the proportion of agreement expected by chance. Kappa has been used as a measure of reproducibility in many epidemiologic settings, such as studies involving twin similarity[13] and control-informant agreement collected from case-control studies. The general expression for the kappa statistic is:

$$k = \frac{P_0 - P_e}{1 - P_e}$$

where, P_0 is the observed proportion of agreement and Pe is the proportion of agreement expected by chance alone[12]. A value of 0 for k indicates no agreement beyond chance and a value of 1 indicates perfect agreement, among many of k's desirable properties[14]. Thus, larger values of k indicate greater agreement between the outcomes.

Here we use k as a measure of agreement between the predicted and observed categorical responses to assess the fit of the GEE model. We estimate k in a second set of estimating equations, similar to Lipsitz[15], Klar[13] and Williamson[16]. With Lipsitz's[10] method, we estimate the probability of the response falling in each of the K categories. Denote this estimated probability for the kth category, tth subunit, of the ith cluster as $\hat{\pi}_{itk}$. We do not have a straightforward predicted response as with linear regression. However, if we did have a predicted response for the tth subunit of the ith cluster, denoted $\hat{Z}_{it}$, it is natural to assume that $\hat{Z}_{it}$ would equal k (k = 1,2,···,K) with probability $\hat{\pi}_{itk}$. Let P_{0it} denote the probability that the predicted response from the model is equal to the observed response, i.e., $\hat{Z}_{it}$= Zit. A natural estimate of P_{0it} is obtained by using $\hat{\pi}_{it}Z_{it}$, the estimated probability from the fitted model that the response falls into the observed category for the *t*th subunit of the *i*th cluster. We define k_{it} as the agreement between the predicted and observed responses for the *t*th subunit of the *i*th cluster as follows:

$$k_{it} = \frac{P_{0it} - P_e}{1 - P_e},$$

where, P_{0it} is defined above and P_e is the probability of correct prediction expected by chance alone.

As an estimate of P_e, we fit an intercept-only model. Cox[17] and Nagelkerke[18] proposed using an intercept-only model as a baseline model when generalizing the coefficient of determination for assessing the fit of a logistic regression model. Thus, we will fit a model with just the intercepts, the λ_k parameters for k = 1,2, ···,K-1, and no covariates. This baseline model will provide a good starting point from which to compare the proposed model. The estimated category probabilities from the intercept-only model will be the same for all clusters and subunits and will be denoted:

$$\hat{p}_{itk} = \hat{p}_k = \sum_{i=1}^{N}\sum_{t=1}^{T} I(Z_{it} = k)/NT = \sum_{i=1}^{N}\sum_{t=1}^{T} Y_{itk}/NT = n_k/NT$$

where, n_k is the sum of observations in category k, i.e., $n_k = \sum_{i=1}^{N}\sum_{t=1}^{T} Y_{itk}$ for k = 1, . . .,K. All n_k observations with response category k will each be correctly predicted with probability $\hat{p}_k$; accordingly, the estimate of P_e is:

$$\hat{p}_e = \hat{p}_k = \sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{t=1}^{T} I(\hat{Z}_{it} = Z_{it} = k)/NT$$

$$= \sum_{k=1}^{K} n_k \hat{p}_k / NT = \sum_{k=1}^{K} \hat{p}_k^2$$

The agreement between two raters for assessing a categorical outcome with K categories can be depicted in a K × K contingency table[14]. The row and column total probabilities for the kth outcome category are p_k and $p_{.k}$, the marginal probabilities that the two raters assess the outcome in the kth category. The estimate of the probability expected by chance is calculated assuming independence between the rows and columns in the contingency table and is $\hat{P}_e = \sum_{k=1}^{K} p_k p_{.k}$, which is similar to the estimate above. We will estimate an overall k to ascertain the fit of the model ($k = k_{it}$ for i = 1, . . .,N and t = 1, . . . , T). By noting that $P_{0it} = P_e+k(1.0-P_e)$, we use a second set of estimating equations as follows. Let P_{0i} and U_i denote the T×1 vectors $\left[P_{0i1},\ldots\ldots,P_{0iT}\right]'$ and $\left[\hat{\pi}_{i1}z_{i1},\ldots\ldots,\hat{\pi}_{iT}z_{iT}\right]'$. The second set of estimating equations are, thus:

$$v_2(k,\beta) = \sum_{i=1}^{N} C_i' W_i^{-1}\left\{U_i(\beta) - P_{0i}(k)\right\} = 0$$

where, $C_i = dP_{0i}/dk = \left[1-\hat{P}_e,\ldots\ldots,1-\hat{P}_e\right]'$ and $W_i \approx \mathrm{var}(U_i)$ is the T ×T working covariance matrix ofU_i. To compute $(\hat{\beta},\hat{k})$, one can use a Fisher-scoring-type algorithm such as:

$$\hat{\beta}^{(m+1)} = \hat{\beta}^{(m)} - \left[\sum_{i=1}^{N}\hat{D}_i'\hat{V}_i^{-1}\hat{D}_i\right]^{-1}\left[\sum_{i=1}^{N}\hat{D}_i'\hat{V}_i^{-1}\left\{Y_i - \hat{\pi}_i\left(\hat{\beta}^{(m)}\right)\right\}\right] \text{ and}$$

$$\hat{k}^{(m+1)} = \hat{k}^{(m)} - \left[\sum_{i=1}^{N}\hat{C}_i'\hat{W}_i^{-1}\hat{C}_i\right]^{-1}$$
$$\left[\sum_{i=1}^{N}\hat{C}_i'\hat{W}_i^{-1}\left\{U\left(\hat{\beta}^{(m+1)}\right) - P_{0i}\left(\hat{k}^{(m)}\right)\right\}\right],$$

where, m denotes the iteration. We use Liang[1]'s empirically corrected variance estimate of $\hat{\beta}$ and Prentice's[19] empirically corrected variance estimate of $\hat{k}$. The second set of estimating equations can be solved non-iteratively if we choose the T × T identity matrix for W_i:

$$\hat{k} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\pi}_{it}z_{it}/NT - \hat{P}_e}{1-\hat{P}_e}$$

The term $\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\pi}_{it}z_{it}/NT - \hat{P}_e$ is the average predicted probability corresponding to the observed responses. If the fitted model predicts the categorical response perfectly, i.e., $\hat{\pi}_{it}z_{it} = 1.0$, then $\hat{k} = 1.0$. If the fitted model predicts the responses no better than an intercept-only model, then $\hat{k} = 0.0$. This kappa-like classification statistic should be interpreted as the average probability of predicting the observed responses above and beyond the prediction by the intercept-only model.

DATA AND VARIABLES

In present study the diabetes mellitus data was used 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. The response variable is defined in terms of the observed glucose level two hours of 75 g glucose load for each follow-up visit. The cut-off point for the blood glucose level is 11.1 mM 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, age, sex, education level, area of residence, family history of father and mother and time in our study. 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. Other variables are dichotomous variable.

RESULTS AND DISCUSSION

The kappa-like statistic takes on a value of 0.0 for the intercept-only model and a value of 1.0 for the saturated model. An advantage of the statistic is that no decisions need be made when calculating it, unlike methods based on covariate partitioning (where to partition, how many partitioned categories), Hosmer and Lemeshow's approach, rank correlation methods and classification tables. Interpretation of the kappa statistic is not always straightforward; see Fleiss[12] and Landis[20] for details. Similar to Landis's[20] labeling of kappa values, Williamson[21] suggest interpreting the values of kappa for this classification index as follows:

Kappa statistic	Fit of model
0.00-0.20	Poor
0.21-0.40	Fair
0.41-0.60	Good
0.61-1.00	Excellent

First we fit a GEE logistic regression model with main effects terms only. We then examined various interaction and quadratic variables for entry into the regression model at the significance level of 0.05. The quadratic terms for age and time were significant and entered into the final model. The kappa-like classification index for this final model increased indicating a better fit. With the inclusion

Table 1: GEE logistic regression models without quadratic terms assuming the various correlation structures for diabetic mellitus study

	Exchangeable			Autoregressive			Pairwise		
Covariates	Estimate	Wald statistic	p-value	Estimate	Wald statistic	p-value	Estimate	Wald statistic	p-value
Intercept	0.3093	0.9891	0.322614	0.3396	1.0466	0.295284	0.2859	0.8421	0.399732
Age	0.0005	0.0942	0.924950	0.0003	0.0669	0.946661	0.0002	0.0538	0.957095
Sex	-0.0860	-0.7189	0.472203	-0.0756	-0.6101	0.541796	-0.0642	-0.5178	0.604598
Edlv	-0.3659	-2.7769	0.005488	-0.3579	-2.6161	0.008894	-0.3359	-2.4352	0.014884
Area	-0.3289	-2.5381	0.011146	-0.3282	-2.4367	0.014814	-0.3025	-2.2071	0.027307
FHFM	0.2071	1.7272	0.084132	0.2006	1.7143	0.086474	0.2011	1.1144	0.265108
Time	0.0705	3.0385	0.002378	0.0795	3.2906	0.000999	0.07001	3.0034	0.002669
Likelihood statistic		179.166			193.368			172.561	
Kappa-like statistic(k)		0.38			0.47			0.18	

Table 2: GEE logistic regression models with quadratic terms assuming the various correlation structures for diabetic mellitus study

	Exchangeable			Autoregressive			Pairwise		
Covariates	Estimate	Wald statistic	p-value	Estimate	Wald statistic	p-value	Estimate	Wald statistic	p-value
Intercept	-6.4615	-0.8831	0.377182	-6.8573	-0.8959	0.370306	-5.1436	-0.7614	0.446418
Age	0.0004	0.0842	0.932897	0.0006	0.0956	0.923838	0.0003	0.0669	0.946661
Sex	-0.1358	-0.8328	0.404958	-0.1556	-0.8713	0.383590	-0.1137	-0.6923	0.488749
Edlv	-0.2561	-1.9776	0.047974	-0.2739	-2.0016	0.045328	-0.2156	-1.7564	0.079020
Area	-0.2872	-2.1751	0.029623	-0.3158	-2.4315	0.015037	-0.2541	-2.0827	0.037279
FHFM	0.1639	1.3392	0.180506	0.1842	1.3027	0.192677	0.1445	1.0415	0.297643
Time	0.0675	2.6138	0.008954	0.0713	2.6732	0.007513	0.0512	2.1497	0.031579
$(Age)^2$	0.0093	2.0954	0.036136	0.0097	2.1305	0.03313	0.0078	2.0002	0.045479
$(Time)^2$	0.4647	3.1867	0.001439	0.4831	3.2717	0.001069	0.3931	3.0026	0.002677
Likelihood statistic		186.647			195.872			173.369	
Kappa-like statistic(k)		0.54			0.58			0.46	

of the two quadratic terms, the model fit the data quite well according to Barnhart's[5] test.

From Table 1, it can be seen that the kappa-like statistic is a fair predictor for the GEE model of exchangeable correlation structure, a good predictor for the autoregressive correlation structure and poor predictor for pairwise correlation structure. Though the likelihood ratio test shows significant effect of covariates in the model but this test does not indicate how well the model predicts the observed responses.

Table 2 shows that when we include the two quadratic terms the kappa-like statistic is good for GEE model of exchangeable correlation structure, a excellent predictor for the autoregressive correlation structure and fair predictor for pairwise correlation structure.

CONCLUSIONS

The kappa-like classification statistic is a more appropriate indicator of how well the model predicts the observed responses at the cluster level (e.g., an individual) as opposed to how well the model fits the data at the group level (e.g., treatment category). Often a model can fit the data well in terms of predicting the proportion of positive responses for a group of individuals, but is not necessarily useful for predicting a particular individual's response. The kappa-like classification index is an intuitive measure for assessing model fit in that it estimates the probability of an observation being correctly predicted by the fitted model. Then, this probability is corrected for chance by comparing it to the probability that an intercept-only model would have correctly predicted the observation. For the diabetes mellitus study, we would recommend that the kappa value for the GEE model with quadratic terms indicated better prediction than the GEE models without quadratic terms. That is, the GEE models with quadratic terms are fitter well.

ACKNOWLEDGMENTS

The authors thank Dr. Kalipada Sen, Professor, Department of Statistics, University of Dhaka, Bangladesh for helpful discussions and manuscript review.

REFERENCES

1. Liang, K.Y. and S.L. Zeger, 1986. Longitudinal data analysis using generalized linear models. Biometrika, 73: 13-22.
2. Zeger, S.L. and K.Y. Liang, 1986. Longitudinal data analysis for discrete and continuous outcomes. Biometrics, 42: 121-130.
3. Horton, N. J., J. D. Bebchuck, C.L. Jones, S.R. Lipsitz, P.J. Catalano, G.E.P. Zahner and G.M. Fitzmaurice, 1999. Goodness-of-fit for GEE: An example with mental health service utilization. Statistics in Medicine, 18: 213-222.

4. Hosmer, D.W. and S. Lemeshow, 1989. Applied Logistic Regression. Wiley, New York.
5. Barnhart, H.X. and J.M. Williamson, 1998. Goodness-of-fit tests for GEE modeling with binary responses. Biometrics, 54: 720-729.
6. Tsiatis, A.A., 1980. A note on a goodness-of-fit test for the logistic regression model. Biometrika, 67: 250-251.
7. Pan, W., 2002. Goodness-of-fit tests for GEE with correlated binary data. Scand. J. Stat., 29: 101-110.
8. Preisser, J.S. and B.F. Qaqish, 1996. Deletion diagnostics for generalized estimating equations. Biometrika, 83: 551-562.
9. Preisser, J.S. and B.F. Qaqish, 1999. Robust regression for clustered data with application to binary responses. Biometrics, 55: 574-579.
10. Lipsitz, S.R., K. Kim and L. Zhao, 1994. Analysis of repeated categorical data using generalized estimating equations. Statistics in Medicine, 13: 1149-1163.
11. Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20: 37-46.
12. Fleiss, J.L., 1971. Measuring nominal scale agreement among many raters. Psychol. Bull., 76: 378-382.
13. Klar, N., S.R. Lipsitz and J. Ibrahim, 2000. An estimating equations approach for modeling kappa. Biometrical J., 42: 45-58.
14. Fleiss, J.L., 1981. Statistical Methods for Rates and Proportions. Wiley, New York.
15. Lipsitz, S.R., N.M. Laird and T.A. Brennan, 1994. Simple moment estimates of the co-efficient and its variance. Applied Stat., 43: 309-323.
16. Williamson, J.M., A.K. Manatunga and S.R. Lipsitz, 2000. Modeling kappa for measuring dependent categorical agreement data. Biostatistics, 1: 191-202.
17. Cox, D.R. and E.J. Snell, 1989. The Analysis of Binary Data. 2nd Edn., Chapman and Hall, London.
18. Nagelkerke, N.J.D., 1991. A note on a general definition of the coefficient of determination. Biometrika, 78: 691-692.
19. Prentice, R.L., 1988. Correlated binary regression with covariates specific to each binary observation. Biometrics, 44: 1033-1048.
20. Landis, R.J. and G.G. Koch, 1977. The measurement of observer agreement for categorical data. Biometrics, 33: 159-174.
21. Williamson, J.N., Hung-Mo Lin and H.X. Barnhart, 2003. A classification statistic for GEE categorical response models. J. Data Sci., 1: 149-165.