Trakya Univ Tip Fak Derg 2007;24(3):222-227

Original Article / ***Klinik Çalışma - Araştırma***

Comparison of Generalized Estimating Equations' Performance in Clustered Binary Observations

Genelleştirilmiş Tahmin Denklemlerinin Performansının Gruplandırılmış İkili Gözlemlerde Karşılaştırılması

Ertuğrul ÇOLAK, Kazım ÖZDAMAR

Department of Biostatistics, Medical Faculty of Osmangazi University, Eskişehir

***Submitted** / **Başvuru tarihi:** 07.06.2007 **Accepted** / **Kabul tarihi:** 19.06.2007*

Objectives: The objective of this study is to compare the performance of generalized estimating equations (GEE) for analysis of clustered binary observations under varying group and observation numbers according to intraclass correlation coefficients (ICC).

Materials and Methods: The comparison of GEE performance was made by using bias of parameter estimations through computer simulations under varying group sizes, ICC, number of clusters, and number of observations per cluster. Simulations were performed in SAS 9.0 by using Monte Carlo simulation method. Analyses were made with SAS GENMOD procedure.

Results: When intraclass correlation coefficient was low (ICC<0.10), there was no significant difference in parameter estimation and their biases and it was observed that GEE gave reliable, consistent and unbiased estimations. However, when ICC increased (ICC>0.10), it was found that the parameter estimations were significantly biased. On the condition that total sample size is fixed; it was observed that, even though the general sample size was constant in all groups while the number of groups was decreasing, when the number of observations per cluster increased, parameter estimations and their biases weren't affected significantly and the effective factor in parameter estimation was ICC.

Conclusion: Because GEE method uses population averaged logistic regression approach, it cannot explain the changes and correlations in clusters completely. The use of GEE method is inconvenient particularly for data sets which have ICC greater than 0.10.

Key Words: Clustered binary observations; intraclass correlation coefficient; generalized estimating equations.

Amaç: Bu çalışmada, gruplandırılmış ikili gözlemler içeren veri setlerinin analizlerinde kullanılan genelleştirilmiş tahmin denklemlerinin (GTD) performansı, farklı grup ve birim sayısında grup içi korelasyon katsayısına (GİKK) göre karşılaştırıldı.

Gereç ve Yöntem: Genelleştirilmiş tahmin denklemleri yönteminin performans karşılaştırmaları, parametre tahminlerinin yanlılıkları kullanılarak farklı grup büyüklükleri, grup içi korelasyon katsayısı, grup sayısı ve her bir gruptaki gözlem sayısında simülasyon çalışmaları yapılarak gerçekleştirildi. Simülasyonlar Monte Carlo simülasyon yöntemi kullanılarak SAS 9.0 programında yapıldı. Analizlerde SAS GENMOD prosedürü kullanıldı.

Bulgular: Grup içi korelasyon katsayısı düşük düzeylerde olduğunda (GİKK<0.10) parametre tahminlerinde ve yanlılıklarında önemli düzeyde bir farklılık gözlenmedi ve GTD yönteminin güvenilir, tutarlı ve yansız tahmin yaptığı saptandı. Ancak GİKK arttıkça (GİKK>0.10), parametre tahminlerinin yüksek oranda yanlı olduğu bulundu. Toplam örnek büyüklüğü sabit kalmak koşuluyla; grup sayısı azalırken genel örnek büyüklüğü tüm gruplarda sabit olmasına rağmen, gruplarda yer alan birim sayıları arttığında parametre tahminlerinin ve yanlılıklarının bu durumdan önemli düzeyde etkilenmediği, parametre tahminlerinde etkili olan faktörün yine GİKK olduğu gözlendi.

Sonuç: Genelleştirilmiş tahmin denklemleri yöntemi popülasyon ortalamalı lojistik regresyon yaklaşımını kullandığından grup içindeki değişimleri ve ilişkiyi iyi düzeyde açıklayamamaktadır. Özellikle GİKK'si 0.10'dan büyük veri setleri için GTD yönteminin kullanılması oldukça sakıncalıdır.

Anahtar Sözcükler: Gruplandırılmış ikili gözlemler; grup içi korelasyon katsayısı; genelleştirilmiş tahmin denklemleri.

Correspondence (İletişim adresi): Dr. Ertuğrul Çolak. Osmangazi Üniversitesi Tıp Fakültesi Biyoistatistik Anabilim Dalı, 26480 Meşelik, Eskişehir. Tel: 0222 - 239 29 79 Fax (Faks): 0222 - 229 01 70 e-mail (e-posta): ecolak@ogu.edu.tr

Clustered binary data arise frequently in medical research and studies. Especially this type of data can be obtained from experimental and observational epidemiologic studies. In experimental studies, such as cross-over trials and stratified cohort studies, clusters are formed from design consideration, while in observational studies, such as twin studies, familial studies and ophthalmologic studies, clusters are formed inherently.[1-3] In cross-over trials, twin and ophthalmologic studies, cluster size is often two. However, cluster size in familial studies can be more than two according to the number of individuals in the family.[2,4,5] Clustered randomized controlled clinical trials where treatments are randomly assigned to clusters, are different type of studies from which clustered data can be obtained.[6-9] Also clustered data are frequently observed in matched case-control studies which are frequently used in health area.[10]

Regardless of study types, if the outcome measurements of interest in clustered data are binary then this type of data are called clustered binary data.[11,12] Common assumption of the studies where clustered data can be obtained is that the observations in the same cluster are independent.[13] On the other hand, this assumption is not hold mostly, because the observations that share the same cluster differ from the other observations characteristically. For that reason, a correlation among the observations in the same cluster is inevitable.[7]

Intraclass correlation coefficient (ICC) is a quantitative measure of correlation or similarity among individuals within clusters and is defined firstly by Fisher.[14-16] The ICC is the measure of variation between and within clusters of individuals and it is an average correlation for the outcome variable obtained from the individuals in the same cluster.[17,18] When the correlated observations are obtained, comparison of the performance of the statistical methods used in the analysis of clustered binary data is very crucial. Generalized estimating equation (GEE) is a statistical method used frequently for analysis of clustered data.[11,12] The objective of this study is to compare the performance of GEE for analysis of clustered binary observations under varying group and observation numbers according to ICC.

MATERIALS AND METHODS

The data sets that contain clustered binary observations consist of K clusters and n_i observations for each cluster. Outcome variable vector for i^{th} cluster is defined as $y'_i=(y_{i1},....y_{in_i})$, $y_{ij}=0.1$ where i=1,....,K, j=1,....,n_i Vector of explanatory variables for j^{th} observation in i^{th} cluster is x'_{ij}= (1, x_{1ij},....,x_{pij}) where p is the number of explanatory variables. X_i= (x_{i1},....,x_{in_i})' is the matrix of explanatory variables for all observations in i^{th} cluster. P (Y_{ij}=1 | x_{ij})=$\pi(x_{ij})$ is the response probability to be modeled for j^{th} observation in i^{th} cluster.

The model, used for clustered binary data, is the population average model that takes the correlation among responses into equations for parameter estimation. Population average (PA) logistic probability model is described as

$$\pi_{PA}(x_{ij})=\frac{\exp(x'_{ij}\beta_{PA})}{1+\exp(x'_{ij}\beta_{PA})}$$

where β_{PA} is the parameter vector. PA model uses average effect of all clusters instead of separate effect of each cluster.[19-21] Some additional notation is required to fully describe the application of GEE to the PA model. Two matrices are used to describe the within-cluster covariance of the correlated observations of the outcome variable. The first is a n_i x n_i diagonal matrix containing the variances under the PA model and denoted

$$A_i=\text{diag}\{\pi_{PA}(x_{ij})x(1-\pi_{PA}(x_{ij}))\}$$

The second is the n_i x n_i exchangeable correlation matrix and denoted

$$R_i(\rho)=\begin{pmatrix}1 & \rho & \dots & \rho\\ \rho & 1 & & \rho\\ \vdots & & \ddots & \vdots\\ \rho & \rho & \dots & 1\end{pmatrix},$$

where

$$(\hat{\rho}) = \frac{1}{(N^{*} - \rho)\phi} \sum_{i=1}^{K} \sum_{j \neq k} e_{ij}\, e_{ik},\ N^{*} = \sum_{i=1}^{K} n_i\ (n_i - 1),$$

$$N = \sum_{i=1}^{K} n_i,\ \phi = \frac{1}{(N - \rho)} \sum_{i=1}^{K} \sum_{j=1}^{n_i} e_{ij}^{2}$$

e_{ij} is the Pearson residual and is defined as

$$e_{ij} = \frac{y_{ij} - \pi_{PA}\,(x_{ij})}{\sqrt{Var\,(\pi_{PA}\,[x_{ij}])}}$$

Also $R_i\,(\rho)$ is known as the working correlation matrix in GEE.[20] Using the fact that the correlation is defined as the covariance divided by the product of the standard deviations it follows that the covariance matrix in the i^{th} cluster is $V_i = A_i^{1/2}\ R_i\,(\rho)\ A_i^{1/2}$. The contribution to the estimating equations for the i^{th} cluster is calculated as $D_i^{'} V_i^{-1} S_i$ where $D_i^{'} = X_i^{'} A_i$. S_i is the vector with j^{th} element the residual $s_{ij} = (y_{ij} - \pi_{PA}\,[x_{ij}])$. The full set of estimating equations is $\sum_{i=1} D_i^{'} V_i^{-1} S_i = 0$. The solution of these equations gives the parameter estimates, $\hat{\beta}_{PA}$.[20, 21]

Comparison of GEE performance in clustered binary observations was made by using Monte Carlo simulation method with 1000 replications. Parameter estimates and their biases were used in comparison. The probability model with one explanatory variable, logit $(\pi_{ij}) = \beta_0 + \beta_1 x_{ij} + e_i$, was used in simulations. The random variable e_i reflects a random effect specific to the i^{th} cluster and the variance of e_i denotes a degree of heterogeneity across the clusters. Here, e_i is assumed stochastically independent of the explanatory variable and normally distributed with mean 0 and unknown variance σ_e^2, $e_i \sim N\,(0, \sigma_e^2)$ Under the probability model, the variance of the random effect e_i, σ_e^2, represents the between cluster variation and the within cluster variation $\pi^2/3$, denotes the variance of logistic distribution where $\pi = 3.14$. From these definitions, ICC can be defined as the ratio of between cluster variation and total variation of the outcome variable and denoted as $ICC = \sigma_e^2 / (\sigma_e^2 + \pi^2/3)$.[12,22]

Simulation algorithm

The probability model with one explanatory variable was used in simulations and the following steps were applied to carry out the simulations.

1) Assume that $\beta_0 = 0$, and set up a value of the parameter β_1.

2) The explanatory variable x_{ij} with $K\ x\ n_i$ size was generated from normal distribution with 0 mean 1 variance.

3) The random effect e_i with K size was generated from normal distribution with 0 mean and σ_e^2 variance.

4) The values obtained from the first three steps were used in probability model and π_{ij} probability values were achieved.

5) The outcome variable y_{ij} was generated from Bernoulli distribution by using the π_{ij} values obtained in step 4, $y_{ij} \sim Bernoulli\ (\pi_{ij})$.

6) GEE analysis was performed by using the outcome variable obtained in step 5 and the explanatory variable achieved in step 2.

7) The parameter estimates were recorded. 5th, 6th and 7th steps were replicated 1000 times. Thus, 1000 different parameter estimates were obtained from the analyses.

In simulation studies, K=20, 100 and 250 were selected for cluster sizes. Three combinations were used for cluster and observations size for each cluster. These combinations are K=250 and n_i=8, K=100 and n_i=20, K=20 and n_i=100. Thus, the total sample size was selected 2000 for all combinations. Seven different values for ICC were selected (ICC=0, 0.05, 0.10, 0.30, 0.50, 0.75 and 0.90). These ICC values were used to calculate σ_e^2 values by using ICC formula. Then, these variances were used in step 3 to generate random variable e_i. The ICC values and corresponding variances are as follows. σ_e^2=0 for ICC=0, σ_e^2=0.17 for ICC=0.05, σ_e^2 =0.37 for ICC=0.10, σ_e^2=1.41 for ICC=0.30, σ_e^2=3.29 for ICC=0.50, σ_e^2=9.87 for ICC=0.75, σ_e^2=29.61 for ICC=0.90. Thus, the outcome variable obtained in step 5 had intraclass correlations depending on the σ_e^2 values.

After the analyses were performed, the mean of the 1000 different parameter estimates was calculated. It was evaluated that how the average of parameter estimates close to the value determined for β_1 in step 1. The biases were calculated by sub-

tracting fix value (0.3) for β_1 from the estimates. Comparisons of GEE performance were made by using these biases. Simulations and analyses were performed by using SAS 9.0 programming language and SAS GENMOD procedure.

RESULTS

The parameter estimates and their biases were displayed in Table 1 according to varying group sizes, ICC, number of clusters, and number of observations per cluster. Biases were calculated as $\hat{\beta}_1 - \beta_1$. Table 1 shows that when ICC<0.10, GEE had ignorable magnitude of bias and GEE gave reliable, consistent and unbiased estimates. However, when ICC increases i.e. ICC>0.10, the parameter estimates obtained from GEE were more biased. On condition that the total sample size is fixed (Kxn_i=2000); when the number of cluster decreases and the number of observations in clusters increases, the parameter estimates were not affected. Thus, it was observed that the most important factor affecting the parameter estimates is ICC. The parameter estimates showed in Table 1 were demonstrated in Fig. 1-3 according to different cluster and observation numbers.

Table 1. Parameter estimates and their biases obtained from 1000 Monte Carlo simulation for $\hat{\beta}_1$=0.3

K	n_i	ICC	$\hat{\beta}_1$	Biases
		0.00	0.2998	-0.0002
		0.05	0.2896	-0.0104
		0.10	0.2859	-0.0141
250	8	0.30	0.2240	-0.0760
		0.50	0.2094	-0.0906
		0.75	0.1893	-0.1107
		0.90	0.1002	-0.1998
		0.00	0.2997	-0.0003
		0.05	0.2971	-0.0029
		0.10	0.2768	-0.0232
100	20	0.30	0.2741	-0.0259
		0.50	0.2001	-0.0999
		0.75	0.1990	-0.1010
		0.90	0.1055	-0.1945
		0.00	0.3014	0.0014
		0.05	0.2870	-0.0130
		0.10	0.2854	-0.0146
20	100	0.30	0.2396	-0.0604
		0.50	0.2350	-0.0650
		0.75	0.1921	-0.1079
		0.90	0.1067	-0.1933

ICC: Intraclass correlation coefficients.

DISCUSSION

It is fact that there is more than one method for analysis of clustered binary data. As a result of

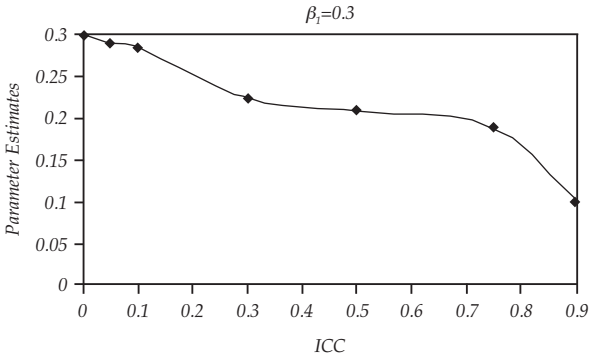

Fig. 1- *Parameter estimates for β_1=0.3, K=250, and n_i=8. ICC: Intraclass correlation coefficients.*

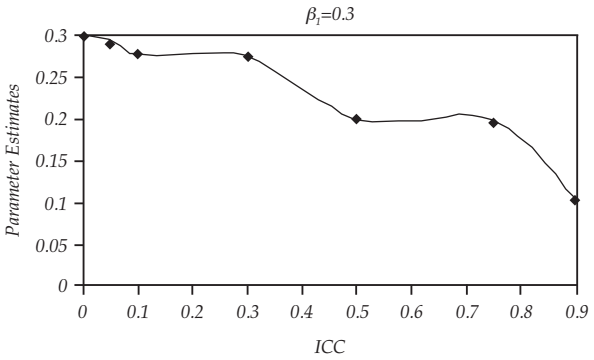

Fig. 2- *Parameter estimates for β_1=0.3, K=100, and n_i=20. ICC: Intraclass correlation coefficients.*

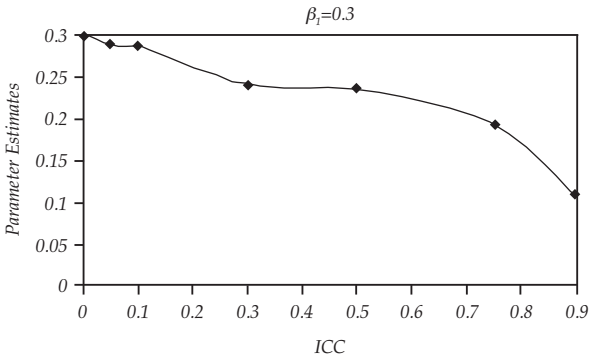

Fig. 3- *Parameter estimates for β_1=0.3, K=20, and n_i=100. ICC: Intraclass correlation coefficients.*

literature researches, it was observed that GEE is in the methods used frequently for analysis of clustered binary data.[2,12] In this study, the performance of GEE in parameter estimation were compared according to ICC. The aim of this comparison is to guide researchers on selection of methods for analysis of clustered binary data. As a result of the simulation, it was observed that GEE gave biased estimates particularly when ICC is greater than 0.10. These results showed the importance of ICC in parameter estimates.

Ananth et al.[11] tried to determine the factors that affect the perinatal mortality (fetal deaths, plus deaths within the first 28 days) with logistic regression methods. They used 285226 twins from 142613 pregnancies (clusters) and showed that ignoring the intraclass correlation in the analyses affects the covariate effects and consequently interpretation of results. Heo and Leon[12] showed that random effect logistic regression method gives more unbiased estimates than GEE when the ICC is high and emphasized that GEE should not be used for clustered binary data having high intraclass correlation. Neuhaus indicated that population averaged approaches of GEE is very popular for analysis of clustered binary data, but GEE does not assess within-subject changes, especially when there is intraclass correlation in the data.[23]

In conclusion, based on the results of this simulation, GEE is preferable for analysis of clustered binary data when there is no intraclass correlation. However, in the analysis of data sets including intraclass correlation, particularly if ICC>0.10, the usage of GEE gives biased, inconsistent and unreliable parameter estimates. Therefore, researchers must determine the intraclass correlation in the data set before the analysis.

REFERENCES

1. Giraudeau B, Mallet A, Chastang C. Case influence on the intraclass correlation coefficient estimate. Biometrics 1996;52:1492-7.
2. Kang W, Lee MS, Lee Y. HGLM versus conditional estimators for the analysis of clustered binary data. Stat Med 2005;24:741-52.
3. Morel JG, Neerchal NK. Clustered binary logistic regression in teratology data using a finite mixture distribution. Stat Med 1997;16:2843-53.
4. Murray DM, Alfano CM, Zbikowski SM, Padgett LS, Robinson LA, Klesges R. Intraclass correlation among measures related to cigarette use by adolescents: estimates from an urban and largely African American cohort. Addict Behav 2002;27:509-27.
5. Neuhaus JM. Estimation efficiency and tests of covariate effects with clustered binary data. Biometrics 1993;49:989-96.
6. Albert JM. Estimating efficacy in clinical trials with clustered binary responses. Stat Med 2002;21:649-61.
7. Reed JF 3rd. Adjusted chi-square statistics: application to clustered binary data in primary care. Ann Fam Med 2004;2:201-3.
8. Smeeth L, Ng ESW. Intraclass correlation coefficients for cluster randomized trials in primary care-data from the MRC trial of the assessment and management of older people in the community. Controlled Clinical Trials 2002;23:409-421.
9. Song JX, Ahn CW. An evaluation of methods for the stratified analysis of clustered binary data in community intervention trials. Stat Med 2003; 22:2205-16.
10. Murray DM, Short BJ. Intraclass correlation among measures related to tobacco use by adolescents: estimates, correlates, and applications in intervention studies. Addict Behav 1997;22:1-12.
11. Ananth CV, Platt RW, Savitz DA. Regression models for clustered binary responses: implications of ignoring the intracluster correlation in an analysis of perinatal mortality in twin gestations. Ann Epidemiol 2005;15:293-301.
12. Heo M, Leon AC. Comparison of statistical methods for analysis of clustered binary observations. Stat Med 2005;24:911-23.
13. Rieger RH, Weinberg CR. Analysis of clustered binary outcomes using within-cluster paired resampling. Biometrics 2002;58:332-41.
14. Ridout MS, Demetrio CGB, Firth D. Estimating intraclass correlation for binary data. Biometrics 1999; 55:137-48.
15. Tian L. Interval estimation and hypothesis testing of intraclass correlation coefficients: the generalized variable approach. Stat Med 2005;24:1745-53.
16. Zou GY, Donner A. Confidence interval estimation of the intraclass correlation coefficient for binary outcome data. Biometrics 2004;60:807-11.
17. Murray DM, Catellier DJ, Hannan PJ, Treuth MS, Stevens J, Schmitz KH, et al. School-level intraclass correlation for physical activity in adolescent girls. Med Sci Sports Exerc 2004;36:876-82.
18. Parker DR, Evangelou E, Eaton CB. Intraclass correlation coefficients for cluster randomized trials in primary care: the cholesterol education and research trial (CEART). Contemp Clin Trials 2005;26:260-7.
19. Molenberghs G, Ryan LM. An exponential family model for clustered multivariate binary data. Environmetrics 1999;10:279-300.
20. Hardin JW, Hilbe JM, (editors). Generalized estimating equations. New York: Chapman &

Hall/CRC; 2003.
21. Hosmer DW. Lemeshow S, (editors). Applied logistic regression. New York: John Wiley & Sons, Inc.; 2000.
22. Patel JK, Kapadia CH, Owen DB, (editors). Handbook of statistical distributions. New York: Marcel Dekker Inc.; 1976.
23. Neuhaus JM. Assessing change with longitudinal and clustered binary data. Annu Rev Public Health 2001;22:115-28.