The Results of Generalized Estimating Equations in the Presence of Monotone Missing Patterns

Monoton Kayıp Veri Kümelerinin Varlığında Genellenmiş Tahmin Denklemlerinin Sonuçları

Bahar TAŞDELEN,[a]
Gülhan OREKİCİ TEMEL[a]

[a]Department of Biostatistics,
Mersin University Faculty of Medicine,
Mersin

Geliş Tarihi/*Received:* 23.05.2012
Kabul Tarihi/*Accepted:* 26.09.2012

Yazışma Adresi/*Correspondence:*
Gülhan OREKİCİ TEMEL
Mersin University Faculty of Medicine,
Department of Biostatistics, Mersin,
TÜRKİYE/TURKEY
gulhan_orekici@hotmail.com

ABSTRACT Objective: During the analysis of ordinal longitudinal datasets, the most frequent problem is missing or incomplete data. In this study, when comparing two independent groups for this type of datasets, the effects of different missing ratio and different sample size on Type I and Type II error rates were investigated. **Material and Methods:** The data for different missing ratio and sample sizes (n=50,100,200,400) using simulation technique were generated. Repeated measurements (at four time points) for each group were generated from multivariate normal distributions using SAS MVN macro and transformed ordinal structure with quintiles method. Completely random monotone missing data for predefined ratios were created. On the comparison of two independent groups using generalized estimating equations (GEE), Type I and Type II error rates were investigated. These simulations were each replicated 1000 times. **Results:** While the Type I error rate was not affected seriously from missing clusters, the Type II error rate was affected. In the complete dataset for both small and large datasets, Type I error rates were <0.0001. While Type II error rates were greater than 0.10 for small sample sizes (n<100), this value was greater than 50% in the presence of missing datasets. **Conclusion:** The sample size and missing ratio are still important tasks for ordinal longitudinal data both in the planning and analyzing stages of an experimental design.

Key Words: Generalized estimating equations; longitudinal; monotone missing pattern; ordinal

ÖZET Amaç: Ordinal yapıdaki boylamsal veri setlerinin analizi sırasında en sık karşılaşılan sorun kayıp veya tamamlanmayan verilerdir. Bu çalışmada, bu tür veri setlerinde bağımsız iki grubu karşılaştırırken farklı örneklem büyüklüğü ve farklı oranlarda kayıp verilerin Tip I hata ve Tip II hata üzerindeki etkileri incelenmiştir. **Gereç ve Yöntemler:** Simülasyon tekniği ile farklı kayıp veri oranları ve örneklem büyüklükleri (n=50,100,200,400) için veri üretilmiştir. Her grupta tekrarlanan (T=4) ölçümler, çok değişkenli normal dağılımdan SAS MVN makrosu kullanılarak üretilmiş, daha sonra çeyreklikler yöntemi kullanılarak ordinal yapıya dönüştürülmüştür. Tamamen tesadüfi olarak, belirlenen oranlarda monoton kayıp veriler oluşturulmuştur. Bağımsız iki grubun genellenmiş tahmin denklemleri (GEE) ile karşılaştırılmasında, Tip I ve Tip II hata oranları incelenmiştir. Her bir simülasyon 1000 kez tekrarlanmıştır. **Bulgular:** Tip I hata oranlarının aksine, Tip II hata oranları kayıp veriden etkileniştir. Küçük örnek genişlikleri için (n<100) Tip II hata oranı 0.10'dan büyük iken, kayıp veri kümelerinin varlığında bu değer 0.50'nin de üstüne çıkmıştır. **Sonuç:** Sıralı yapıdaki longitudinal verilerden oluşan verilerin planlanması ve analizinde örneklem büyüklüğü ve kayıp verilerin varlığı hala önemlidir.

Anahtar Kelimeler: Genellenmiş tahmin denklemleri; boylamsal; monoton kayıp yapısı; sıralı

Turkiye Klinikleri J Biostat 2013;5(1):1-7

Longitudinal studies are frequently undertaken in medical sciences, and investigators suffer from the presence of incomplete data. An adverse event, the disease state, the lack of efficacy of the study treatment, length of the trial or the refusal of the subject to continue the study cause dropouts.[1,2] In some cases, patients remain on the study, but they do not receive the treatment as planned in the study protocol. To deal with this problem, the intention-to-treat (ITT) principle is often invoked. According to basic ITT principle, patients in the trials should be analyzed in the groups to which they are randomized, regardless of the appropriateness of their eligibility determination, of the treatment they actually received, or of deviation from the protocol.[3]

The terminology of Little and Rubin is used to classify missingness mechanisms. A nonresponse process is missing completely at random (MCAR) if the missingness is independent of both unobserved and observed data and missing at random (MAR) if depends only on the observed data. A process that is neither MCAR nor MAR is termed nonrandom (MNAR). MCAR and MAR are ignorable if the parameters that describe the measurement process are functionally independent of the parameters that describe the missingness process.[4] In monotone missing patterns which are common in medical studies, if a variable Y_j $(j=1,...,p)$ is missing for a particular individual, all subsequent variables Y_k $(k > j)$ are missing for that individual (Table 1).

Multivariate analysis of variance (MANOVA) for repeated measurements is frequently used to analyze continuous longitudinal data. This method has some disadvantages such as exclusion of subjects who have missing repeated measurements. MANOVA is a complete case analysis and is only valid when the missingness mechanism is MCAR. For missing longitudinal data, some imputation methods such as last observation carried to forward (LOCF), hot deck imputations such as similar response pattern imputation, k-NN (k*th* nearest neighboring) imputation and multiple imputation are implemented. Although LOCF is computationally simple and used frequently in clinical trials because of consistency with the intention-to-treat principle, it may give biased estimates of treatment effects and standard errors.[3] For continuous outcomes, the use of a likelihood-based ignorable analysis such as general linear mixed-effects model is suggested and MAR mechanism is required.[4-6]

TABLE 1: Monotone missing data patterns.

Pattern	Y1	Y2	Y3	Y4
Type 1	X	X	X	missing
Type 2	X	X	missing	missing
Type 3	X	missing	missing	missing

For categorical and discrete outcomes (counts), generalized linear mixed models (GLMM) and generalized estimating equations (GEE) were discussed and implemented in the presence of dropouts.[1] The subjects with incomplete data are not excluded from the analysis when using GEE method.[7] However, if the sample size is very small and the missing data mechanism is not MCAR, GEE results can be biased and inconsistent.[8-11] Therefore, weighted GEE approach is proposed to estimate parameters under MAR assumption.[12,13] In the multivariate framework, sample-size calculations for GEE analysis are proposed for binary, count, ordinal and continuous responses.[14-20]

In this simulation study, ordinal longitudinal data were considered and monotone missing patterns were created. Then, the effects of different missing ratio and different sample size on Type I and Type II error rates for the comparison of two independent groups were investigated.

MATERIAL AND METHODS

According to Lipsitz et al.,[18] let assume N independent clusters (i=1,…, N) with T_i repeated responses $(Y_1,\ldots,Y_T)$ for the *i*th cluster $(t=1,\ldots,T_i)$. Let $Z_{it}=k$, $(k=1,\ldots,K)$ denote the ordinal response for the *t*th measurement of the *i*th cluster. For the ease of random sampling, a monotone missing data pattern is assumed here. Furthermore, the study population is assumed as homogeneous and additional covariate terms except group effect are not used. Let $Y_{itk}=I(Z_{it}=k)$ be a binary random variable that takes a value of 1 if the response for the *t*th measurement of the *i*th cluster is in the *k*th category. Then, the marginal probabilities of Z_{it} for all categories are denoted by

$$\Pr[Z_{it}=k]=E[Y_{itk}]=\Pr[Y_{itk}=1]=\pi_{itk} \qquad (1)$$

the corresponding marginal cumulative probabilities are denoted by

$$\Pr[Z_{it} \le k] = \vartheta_{itk} \tag{2}$$

(i=1,…,N ; t =1,…, T ; and k =1,…, K).[21]

The correlated multinomial response data can be analyzed by using following link function of the cumulative marginal response probabilities,

$$h(x) = \log\left(\frac{\Pr[Z_{it} \le k]}{1 - \Pr[Z_{it} \le k]}\right) = \gamma_k + \mathbf{\kappa}^T \mathbf{x}_{it} \tag{3}$$

where γ_k is the *k*th intercept, κ is the p×1 parameter vector, and x_{it} is the covariate vector for the *t*th measurement of the *i*th cluster. For ordinal data, link function can be either cumulative logistic or cumulative probit functions.[18-21] By using logistic function,

$$p_{it} = \Pr[Z_{it} \le k] = \frac{\exp(\gamma_k + \mathbf{\kappa}^T \mathbf{x}_{it})}{1 + \exp(\gamma_k + \mathbf{\kappa}^T \mathbf{x}_{it})} \tag{4}$$

Because the primary objective is to compare groups A and B (for example, treatment and placebo), $p_{AT} = (p_{iT}|X_i = 1)$ and $p_{BT} = (p_{iT}|X_i = 0)$ probabilities are estimated. For this purpose, maximum likelihood estimates of the regression coefficients $\beta = (\gamma, \kappa)^T$ are obtained by an iterative weighted least squares method.[22] In the GEE approach, the score equations take the form

$$\sum [D_i]^T (V_i)^{-1} \{Y_i - p_i(\beta)\} = 0, \tag{5}$$

where D is a vector of partial derivatives $\partial p_i / \partial \beta$ and V_i is a working covariance matrix for the observed data Y_i

$$V_i = A_i R_i A_i, \tag{6}$$

where $A_i = Diag\{\sqrt{p_{it}(1 - p_{it}}\}$ and R_i is a working correlation matrix.[22] Although there are some possible forms of R_i such as identity, exchangeable, and AR-1 autoregressive, unstructured working correlation matrix is the most common used in practice.[8]

SIMULATIONS

In this study, two separate simulation algorithms were considered to compare treatment groups using GEE when the longitudinal data set included incomplete measurements and the effects of different missing ratios and different sample sizes (n=50,100,200,400) on Type I and Type II error rates were investigated (Figure 1). In each simulation, clusters were considered as four-repeated (*T*=4) measurements and the data were generated from multivariate normal distribution using SAS MVN macro (downloaded from web site www.support.sas.com) with 1000 replications.

Step 1. Firstly, assuming no statistically significant difference between treatment means (m1-m2=0). The mean vector for each treatment group was set as $\overline{X}_i = [X_{i1} \quad X_{i2} \quad X_{i3} \quad X_{i4}] = [10 \quad 20 \quad 30 \quad 40]$ (i=1,2). The multivariate normal distributed data

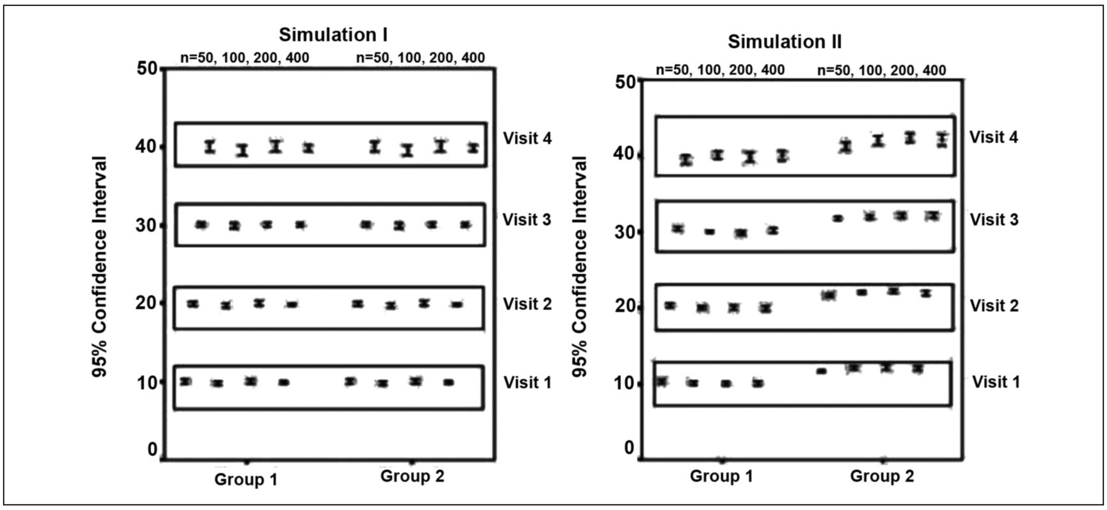

FIGURE 1: Mean values and confidence intervals for simulation by treatment and visits.

were generated with AR-1 autoregressive covariance matrix given in following way. Where σ value was assumed as 2, maximum coefficient of variation was 20% for within measurement homogeneity (for example, CV_1=2/10=20%, CV_2=2/20= 10%, CV_3=2/30=7%, CV_4=2/40=5%), for all groups. Also, high correlations among repeated measurements were assumed.

$$\sigma^2 \begin{bmatrix} 1 & 0.8 & 0.8^2 & 0.8^3 \\ 0.8 & 1 & 0.8 & 0.8^2 \\ 0.8^2 & 0.8 & 1 & 0.8 \\ 0.8^3 & 0.8^2 & 0.8 & 1 \end{bmatrix}$$

Step 2. In the second simulation algorithm, by considering statistically significant minimum effect size, the mean vectors for two treatment groups were $\bar{X}_1 = [10 \quad 20 \quad 30 \quad 40]$ and $\bar{X}_2 = [12 \quad 22 \quad 32 \quad 42]$ separately. The same covariance matrix was also used.

Step 3. For both the first and the second studies, the ordinal responses (*K*=5 categories) were created using 20th, 40th, 60th, and 80th percentiles of normal distribution.

Step 4. After the complete longitudinal data had been generated, monotone missing patterns were created by deleting post-baseline responses with different ratios (p: 25%, 50% 75%) (Table 2). In the datasets, deleted values of dependent variables were chosen in a monotone missing completely at random manner. Monotone missing patterns were created so that missing ratio for the patterns included one missing observation was relatively greater than the others. Any explanatory variable was not included in the longitudinal datasets.

Step 5. All statistical analyses were carried out on both complete and incomplete datasets and GENMOD procedure was used to generate the GEE results. When the means of two treatment groups were similar, the number of rejected H_0 hypothesis (H_0: $\mu_1 = \mu_2$) was declared as the Type I error probability. In addition, when the means of two treatment groups were different with minimum effect size (Δ=2), the number of accepted H_0 hypothesis was declared as the Type II error probability.

TABLE 2: The proportions of deleted post-baseline units to monotone missing patterns.

Sample size	Missing	Monotone missing pattern		
per group	ratio	Type 1 (%)	Type 2 (%)	Type 3 (%)
50	25%	1.00	5.00	19.00
	50%	2.00	11.00	37.00
	75%	3.00	15.00	57.00
100	25%	0.50	4.00	20.50
	50%	1.50	8.50	40.00
	75%	2.00	12.50	60.50
200	25%	0.50	4.50	20.00
	50%	1.25	8.75	40.00
	75%	1.75	13.25	60.00
400	25%	0.63	3.87	20.50
	50%	1.37	8.13	40.50
	75%	1.87	14.50	58.63

Step 6. The estimates of regression coefficients were averaged.

RESULTS

The first simulation results indicated that the type I error rate increased as the proportion of missing clusters increased (π>50%), for all sample sizes (Table 3). Furthermore, parameter estimates and their standard deviations seriously increased as the proportion of missing clusters increased. But similarity between two groups remained. In the complete dataset for small sample size (n=50 per group), parameter estimate and odds ratio were 0.0008±0.0538 and 1.0007±0.0677, separately. The difference between treatment groups was not statistically significant. In the dataset which the ratio of missing cluster was 75%, parameter estimate and odds ratio were 0.0575±0.6401 and 1.0297± 0.3844, separately. The difference between treatment groups was still not statistically significant.

In the first simulation which sample sizes were greater than or equal to 100, the risky group changed for some missing proportions. When sample size was 400, the risky group was the first group for the complete dataset. On the other hand, the risky group was the second group for all incomplete data sets.

While the Type I error rate was not effected seriously from missing clusters, the Type II error

TABLE 3: Regression coefficients (their standard errors) and OR values (their standard errors) derived from GEE analysis and Type I error rates (α).

	Complete dataset			25% missing			50% missing			75% missing		
Sample size per group	**Estimate**	**OR**	α	**Estimate**	**OR**	α	**Estimate**	**OR**	α	**Estimate**	**OR**	α
n=50	0.0008	1.0007	0.000	0.0057	1.0057	0.000	0.0074	1.0074	0.006	0.0575	1.0592	0.020
	(0.0538)	(0.0677)		(0.2338)	(0.1545)		(0.3587)	(0.2213)		(0.6401)	(0.3844)	
n=100	-0.0009	0.9991	0.000	0.0089	1.0089	0.000	0.0130	1.0131	0.004	0.0082	1.0082	0.020
	(0.0229)	(0.0217)		(0.1516)	(0.1097)		(0.2427)	(0.1651)		(0.4354)	(0.2323)	
n=200	-0.0007	0.9993	0.000	-0.0055	0.9945	0.000	0.0040	1.0040	0.008	0.0100	1.0100	0.022
	(0.0155)	(0.0271)		(0.1117)	(0.0734)		(0.1782)	(0.1130)		(0.3148)	(0.1822)	
n=400	-0.0001	0.9999	0.000	0.0006	1.0006	0.000	0.0075	1.0075	0.004	0.0039	1.0039	0.024
	(0.0146)	(0.01288)		(0.0726)	(0.0476)		(0.1271)	(0.0813)		(0.2055)	(0.1274)	

TABLE 4: Regression coefficients (their standard errors) and OR values (their standard errors) derived from GEE analysis and Type II error rates (β).

	Complete dataset			25% missing			50% missing			75% missing		
Sample size per group	**Estimate**	**OR**	β	**Estimate**	**OR**	β	**Estimate**	**OR**	β	**Estimate**	**OR**	β
n=50	1.8257	6.2071	0.000	1.8462	6.3357	0.000	1.8771	6.5345	0.034	1.8756	6.5247	0.096
	(0.0114)	(0.0123)		(0.0170)	(0.0165)		(0.0227)	(0.0175)		(0.0358)	(0.0223)	
n=100	1.7925	6.0044	0.000	1.8189	6.1651	0.000	1.8373	6.2795	0.000	1.9081	6.7403	0.082
	(0.0077)	(0.0082)		(0.0103)	(0.0096)		(0.0152)	(0.0120)		(0.0267)	(0.0173)	
n=200	1.7828	5.9464	0.000	1.7999	6.0490	0.000	1.8334	6.2551	0.000	1.8326	6.2501	0.004
	(0.0058)	(0.0061)		(0.0083)	(0.0073)		(0.0118)	(0.0091)		(0.0181)	(0.0122)	
n=400	1.7735	5.8914	0.000	1.7954	6.0218	0.000	1.8137	6.1331	0.000	1.8260	6.2090	0.000
	(0.0038)	(0.0043)		(0.0055)	(0.0049)		(0.0075)	(0.0058)		(0.0119)	(0.0076)	

rate was effected. In the complete dataset for both small and large datasets, Type I error rates were less than 0.0001. The Type II error rate was greater than 0.10 for small sample sizes, when the proportion of missing clusters was greater than 50% (Table 4).

DISCUSSION AND CONCLUSION

When the studies evaluating GEE analysis performance have been investigated, we have not come up with a study in which monotone loss affect the power of the test and Type 1 error. However, especially in longitudinal clinical trials, monotone missing patterns are frequently observed. For example, in a clinical trial study in which the strength of pain is evaluated, it is possible not to reach the patients that had been observed before. In these cases, in point of changes occuring independent variable in a period, performance evaluation of methods that can be used to compare two groups is important to shed light on studies in the future.

In this paper we have examined the consequences of monotone missing data in longitudinal studies for the results of parameter estimates, the Type I and Type II error rates. Although some researchers have been said that the GEE results could be biased and inconsistent if the missing data mechanism was not MCAR, according to results of simulations monotone missingness caused especially important changes in Type 2 error rates. The losses that were greater than or equal to 50% caused the changes in Type 1 error and Type 2. There was a change in the way of prediction and the unrealistic OR values were obtained. There was also a great increase in standard deviations of the predictions. In this study it was shown that in longitudinal studies including monotone missing completely at random observations, missing ratio should not be exceed 50% to obtain realistic OR values and standard deviations.

Although monotone missing patterns are not important for relatively short-term study (T=4), they may be more significant for long-term studies (T>4). Therefore, the sample size and missing ratio are still important tasks for ordinal longitudinal data both in the planning and analyzing stages of an experimental design.

REFERENCES

1. Minini P, Chavance M. Sensitivity analysis of longitudinal binary data with non-monotone missing values. Biostatistics 2004;5(4):531-44.
2. Mallinckrodt CH, Kaiser CJ, Watkin JG, Detke MJ, Molenberghs G, Carroll RJ. Type I error rates from likelihood-based repeated measures analyses of incomplete longitudinal data. Pharmaceut Statist 2004;3(3):171-86.
3. Beunckens C, Molenberghs G, Kenward MG. Direct likelihood analysis versus simple forms of imputation for missing data in randomized clinical trials. Clin Trials 2005;2(5):379-86.
4. Jansen I, Beunckens C, Molenberghs G, Verbeke G, Mallinckrodt C. Analyzing incomplete discrete longitudinal clinical trial data. Statistical Science 2006;21(1):52-69.
5. Verbeke G, Molenberghs G. Selection Models. Linear Mixed Models for Longitudinal Data. 2nd ed. New York: Springer; 2000. p. 257-62.
6. Molenberghs G, Thijs H, Jansen I, Beunckens C, Kenward MG, Mallinckrodt C, et al. Analyzing incomplete longitudinal clinical trial data. Biostatistics 2004;5(3):445-64.
7. Twisk J, de Vente W. Attrition in longitudinal studies. How to deal with missing data. J Clin Epidemiol 2002;55(4):329-37.
8. Li X, Mehrotra DV, Barnard J. Analysis of incomplete longitudinal binary data using multiple imputation. Stat Med 2006;25(12): 2107-24.
9. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73(1):12-22.
10. Paik MC. Repeated measurement analysis for nonnormal data in small samples. Commun Statist Simul Comp 1988;17(4):1155-71.
11. Laird NM. Missing data in longitudinal studies. Stat Med 1988;7(1-2):305-15.
12. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association 1995;90(429):106-21.
13. Fitzmaurice GM, Molenberghs G, Lipsitz SR. Regression models for longitudinal binary responses with informative drop-outs. Journal of the Royal Statistical Society Series B 1995;57(4):691-704.
14. Liu G, Liang KY. Sample size calculations for studies with correlated observations. Biometrics 1997;53(3):937-47.
15. Rochon J. Application of GEE procedures for sample size calculations in repeated measures experiments. Stat Med 1998;17(14): 1643-58.
16. Pan W. Sample size and power calculations with correlated binary data. Control Clin Trials 2001;22(3):211-27.
17. Liu A, Shih WJ, Gehan E. Sample size and power determination for clustered repeated measurements. Stat Med 2002;21(12):1787-801.

18. Lipsitz SR, Fitzmaurice GM. Sample size for repeated measures studies with binary responses. Stat Med 1994;13(12):1233-9.

19. Shih WJ. [Sample size and power calculations for periodontal and other studies with clustered samples using the method of generalized estimating equations]. Biometrical Journal 1997;39(8):899-908.

20. Jung SH, Ahn C. Sample size estimation for GEE method for comparing slopes in repeated measurements data. Stat Med 2003;22(8): 1305-15.

21. Kim HY, Williamson JM, Lyles CM. Sample-size calculations for studies with correlated ordinal outcomes. Stat Med 2005;24(19): 2977-87.

22. Armitage P, Colton T. Generalized Estimating Equations. Encyclopedia of Biostatistics. 2nd ed. Chichester, West Sussex: John Wiley; 2005. p. 2074-84.