Araştırma Makalesi / Research Article

Use of Generalized Estimating Equations with Multiple Imputation for Missing Longitudinal Data

Gazel Ser[*], Hayrettin Okut

Van Yuzuncu Yil University, Faculty of Agriculture, Department of Animal Science, Van/Turkey
[*] e-mail: gazelser@yyu.edu.tr

Abstract: This study aimed to assess the performance of multiple imputation for the Generalized Estimating Equation (GEE) method. Observations with longitudinal data structure obtained from 1044 individuals during the five years were used. Smoking frequency, response variable with Poisson distribution and the independent variables thought likely to affect these were taken into consideration. Four different working correlation structures were examined to determine the study correlation structure in GEE. Quasi information criterion was used to determine the most appropriate working correlation structure to fit the data. In estimating the missing observation, the missing observations were assumed to be missing at random, and missing observations were estimated using multiple imputation (MI). Thus, the GEE method was applied again to the complete data set obtained and MI-GEE results were obtained. As a result, the appropriate working correlation structure for missing-GEE and MI-GEE was determined as the independent structure, and parameter estimations were obtained using this structure. In both cases, empirical standard error results were evaluated. Accordingly, in the data set with missing observations, effect of alcohol use and family relationship status (p<0.001) and of age (p<0.01) on smoking was found to be significant in missing-GEE results. In the MI-GEE results, effect of intercept, alcohol use, family relationship score, gender, age (P<0.001) and the score of the individual’s relationship with his/her family (p<0.01) was found to be significant. Also, standard error estimations obtained for MI-GEE were much smaller according to the missing-GEE.

Key words: Marginal models, working correlation, multiple imputation

Eksik Uzun Süreli Veride Çoklu Atama ile Genelleştirilmiş Tahmin Eşitliklerinin Kullanımı

Özet: Bu çalışmada, Genelleştirilmiş Tahmin Eşitlikleri (GTD) yöntemi için çoklu atamanın performansının değerlendirilmesi amaçlamıştır. Çalışmada, beş yıl süresince 1044 bireyden elde edilen uzun süreli veri yapısına sahip gözlemler kullanılmıştır. Sigara kullanım sıklığı Poisson dağılışına sahip cevap değişkeni ve bunun üzerine etkili olabileceği düşünülen değişkenler göz önünde bulundurulmuştur. GTD’de çalışma korelasyon yapısının belirlenmesinde dört farklı korelasyon yapısı incelenmiştir. Veri yapısına en uygun çalışma korelasyon yapısının belirlenmesinde yarı olabilirlik bilgi ölçütü kullanılmıştır. Eksik gözlemlerin tahmininde, eksik gözlemlerin şansa bağlı olduğu kabul edilmiş ve çoklu atama (MI) yöntemi uygulanarak eksik gözlemler tahmin edilmiştir. Böylece elde edilen tam veri setine tekrar GTD yöntemi uygulanarak MI-GTD sonuçları elde edilmiştir. Sonuç olarak, eksik veri için GTD ve MI-GTD için en uygun korelasyon yapısı bağımsız yapı olarak belirlenmiş ve bu yapı kullanılarak parametre tahminleri elde edilmiştir. Her iki durumda da ampirik standart hata tahmin sonuçları değerlendirilmiştir. Buna göre, eksik gözleme sahip veri setinde GTD sonuçları ve ampirik standart hata tahminlerinde sigara kullanımı üzerine alkol kullanımının ve aile ilişki sokurunun (p<0.001) ve yaşın etkisi (p<0.01) önemli bulunurken, MI-GTD sonuçlarında ise intersept, alkol kullanımı, aile ilişki skoru, cinsiyet, yaş (P<0.001) ve bireyin ailesiyle olan ilişki skoru (p<0.01) önemli bulunmuştur. Ayrıca, MI-GTD’deki standart hata tahminleri, GTD’ye göre daha küçük elde edilmiştir.

Anahtar kelimeler: Marjinal modeller, çalışma korelasyon, çoklu atama

Introduction

In longitudinal studies, the decision about which analysis method is to be used is based on the structure and distribution of the dependent variable. While marginal models, and the Generalized Linear Model (GLM) and Generalized Estimation Equation (GEE) used as extensions of these models are employed in cases where the dependent variable does not have normal distribution; in subject-specific models, the Generalized Linear Mixed Model (GLMM) is employed (Singer and Willet 2003; O'Brien and Fitzmaurice 2005; Antonio and Beirlant 2007). Marginal models or population-average models used in cases where the dependent variable does not exhibit normal distribution define the distribution possessed by random variables in an exponential distribution family. Then a link function is used between the elements on both sides of the equation. Moreover, the GEE in this model class is not concerned with the joint distribution of the dependent variable, instead it uses marginal distribution of the repeated measurements in the time range. The difference between GEE compared to other estimation methods is that it takes the structure of correlation between repeated measurements into account (Zeger *et al.* 1988; Fitzmaurice and Verbeke 2009). The most important advantage of GEE is that it provides parameter estimates and their asymptotically correct standard errors, confidence intervals, and significance tests, even if the correlation structure is not correctly defined (Pekar and Brabec 2018). GEE allows flexibility in the modeling and interpretation of situations where observations of measurement values are non-normally distributed data (e.g., binary, Poisson, etc.) (Jiang, 2007). Marginal models are formulated simpler than alternative approaches in the analysis of correlated data and provide a more direct approach (Pekar and Brabec 2018).

Material and Methods

The research material used 1044 individual observations obtained during 5 years. The study considered the response variable, where the frequency of smoking of individuals exhibits Poisson distribution, and the variables thought to affect these. These variables are the individual's alcohol use frequency (IAUF), the score for friend influence on individual's smoking (SFIS), the score for individual's listening to his/her family (FLIS), individual relationship score (IRS), family relationship score (FRS), marriage status of parents (PMS), gender and age. In the study, analyses were conducted in two phases in the missing observation longitudinal dataset. In the first phase, marginal model analysis results were derived in the missing observation longitudinal dataset. In the second phase, missing observations were assumed to be MAR, and missing observations were estimated with the Multiple Imputation (MI) method using the MCMC technique. The marginal model was similarly reapplied to the dataset in which missing observations were estimated. In order to determine the structure of correlation between observations in GEE, four different correlation structures were examined. The MI method was used to estimate missing observations. Analysis results were obtained with the MCMC technique of the MI method. The PROC MI in the SAS 9.2 software program was used for analysis. The study used the GEE method for marginal model analyses. The GEE analysis was completed with the PROC GENMOD procedure in the SAS 9.2 (SAS, 2010) statistical software program.

Marginal models

Marginal models for longitudinal data is generated as follows:

$$g^{-1}\left(\mu_{ij}\right)=\eta_{ij}=X_{ij}\beta \qquad (1)$$

The conditional expected value of each response is $(E\left(Y_{ij} \mid X_{ij}\right)=\mu_{ij})$, and a link is established for both sides of the equation with a known link function $(g^{-1}(.))$. In Equation 1, (β) is the $p \times 1$ regression vector (Lipsitz and Fitzmaurice 2009).

Let us assume the longitudinal response variables derived have Poisson distribution. Accordingly, the marginal model is generated as follows:

$$\log(\mu_{ij}) = \eta_{ij} = X_{ij}\beta \qquad (2)$$

In the equation, average of the response variable is linked to the covariance by the *log* link function. Different correlation structures are used in determination of intra-individual relationships between repeated responses (Hardin and Hilbe 2003; Fitzmaurice *et al.* 2004; Lee and Nelder 2004).

Parameter estimations for marginal models

In marginal model models, the GEE model is used for parameter estimation. GEE is a quasi-likelihood-based method, through which estimation equations are derived without the need to exactly define joint distribution. Instead, only likelihood is defined for marginal distributions, and a working correlation matrix $R_i(\alpha)$ is defined for the vector of repeated measurements derived from each individual. $R_i(\alpha)$ is calculated for repeated measurements of each individual (Barnett et al. 2010). Estimation of β in GEE is found with the following equation:

$$S(\beta) = \sum_{i=1}^{K} \frac{\partial \mu_i^{'}}{\partial \beta} V_i^{-1}(Y_i - \mu_i(\beta)) = 0 \qquad (3)$$

In the equation, $\mu_i = (\mu_{i1}, \ldots\ldots, \mu_{it_i})$ is the averages vector; $Y_i = (Y_{i1}, \ldots\ldots, Y_{it_i})$ is the observation vector and V_i is an estimator of covariance matrix of Y_i. These equations defined are similar to those in GLM. The difference is that the averages vector is used instead of a single average, and a covariance matrix is used instead of scalar variance (Liang and Zeger 1986; Li *et al.* 1998; Yeşilova *et al.* 2006).

Since more than one observation were made of the same individual at different time points, a relationship is available between the observations, and this relationship is included in the model as the covariance matrix. In GEE, this covariance matrix is defined as “working covariance”. In equation 3, V_i is the working covariance matrix derived for the observation values Y_is, and is shown as follows:

$$V_i = \phi A_i^{1/2} R_i(\alpha) A_i^{1/2} \qquad (4)$$

In equation 4, A_i is a $t_i \times t_i$ diagonal matrix. For each $y_i = (y_{i1}, \ldots\ldots, y_{it_i})$, $t_i \times t_i$ working correlation matrix $(R_i(\alpha))$ is calculated. If $t_i = 1$, GEE equals GLM (Johnson, 1996).

Several approaches have been suggested by Liang and Zeger (1986) for the covariance parameter to find the estimation values of (β) in determination of the correlation matrix. These are empirical and model-based approaches. Model-based estimation of $Cov(\hat{\beta})$ is as follows:

$$Cov(\hat{\beta}) = \left[\sum_{i=1}^{K} D_i^{'} V_i^{-1} D_i\right]^{-1} \qquad (5)$$

A consistent estimation of the covariance matrix of $(\hat{\beta})$ is derived once the model and correlation matrices are determined accurately. Empirical (sandwich, robust) estimator of $Cov(\hat{\beta})$ is as follows:

$$Cov(\hat{\beta}) = \left(\sum_{i=1}^{K} D_i^{'} V_i^{-1} D_i\right)^{-1} \left\{\sum_{i=1}^{K} D_i^{'} V_i^{-1} (y_i - \mu_i)(y_i - \mu_i)^{'} V_i^{-1} D_i\right\} \left(\sum_{i=1}^{K} D_i^{'} V_i^{-1} D_i\right)^{-1} \qquad (6)$$

Even if the correlation matrix is determined to be wrong, the estimation of the covariance matrix of $(\hat{\beta})$

maintains its consistency. Therefore, the empirical covariance matrix is preferred for applications (Park *et al.* 1998; Aktaş, 2005; Halekoh, 2008). In the GEE method, correlation structure needs to be known to estimate the coefficients concerning the variables. Therefore, the most accurate relationship should be determined. Since the GEE method does not require distribution assumption and is not quasi-likelihood-based, the Quasi Likelihood under Independence Criterion (QIC) was developed (Hin and Wang 2009; Tan *et al.* 2009).

Multiple imputation method

The MI method includes the Monte Carlo technique using derived versions in place of a certain number of missing measurements (m) with (m>1) (Baygül, 2007). In this technique random variables are pulled from probability distributions with the aid of Markov chains. A Markov chain is a sequence formed of random variables, and is a method based on each individual value in the distribution being linked to the previous value in the sequence. The MCMC technique creates as chain with sufficient length for the relevant distribution and ensures stability of the distribution. Alternatively, random variables may be derived by creating more than one chain with sufficient length (Schaffer, 1999). The MI method requires three basic steps; the first is imputation, the second is analysis and the third is aggregation. Parameters are estimated from data clusters collated in the imputation stage. The analysis of the imputed data in the repeated analysis stage is simpler than analysis before imputation because the problem of missing measurements has been resolved. The aggregation stage involves the calculation of (*p*) values, confidence intervals, variance and means for (*m*) times repeated analyses. The values imputed with the MI method should preserve the original structure of the data set (Allison, 2000; Buuren, 2007).

Result and Discussion

As the smoking habit of individuals displayed Poisson distribution, GEE was used for parameter estimation in marginal or population average models. A completed dataset was obtained by estimating the missing observations with the Multiple Imputation (MI) method. Results were obtained by applying the GEE method to this dataset. Table 1 presents the Quasi Information Criterion results for the missing observation state (GEE) and the estimation of missing observations with the MI method (MI-GEE). When the correlation structures given in Table 1 are assessed, the smallest QIC for the GEE and MI-GEE results was obtained for the independent correlation structure. As seen in the table, it is noteworthy that the imputations with the MI method preserved the correlative structure of the data set (see Table 1).

In both situations, all observations obtained from individuals over time were independent of each other. According to the independent correlation structure, the “working” correlation matrix $(R(\alpha))$ is given in Table 2. Missing data-GEE and Complete Data MI-GEE results are given in Table 3

Table 1. Results concerning correlation structure in GEE and MI-GEE methods

Correlation Structure	**Missing Data -GEE- QIC**	**MI-GEE-QIC**
Exchangeable	-1713.9343	-11507.4396
Independent	-1730.1110	-11666.3158
First Order Autoregressive (AR(1)	-1699.2917	-11290.3144
Unstructured (UN)	-1697.0981	-11251.4566

Table 2. “Working” correlation matrix in the independent correlation structure

	year=1	year=2	year=3	year=4	year=5
year=1	1.0000				
year=2	0.0000	1.0000			
year=3	0.0000	0.0000	1.0000		
year=4	0.0000	0.0000	0.0000	1.0000	
year=5	0.0000	0.0000	0.0000	0.0000	1.0000

Table 3. GEE and MI-GEE estimation results and empirical standard error estimations according to independent correlation structure

	Missing Data-GEE		**Complete Data-MI-GEE**	
Parameters*	**Estimation (Std. Dev.)**	**Z**	**Estimation (Std. Dev.)**	**Z**
Intercept	-0.326(0.305)	-1.07	-0.677 (0.116)	-5.85^{***}
IAUF	0.163 (0.012)	13.29^{***}	0.051 (0.003)	18.94^{***}
SFIS	-0.029(0.023)	-1.29	-0.011 (0.009)	-1.26
FLIS	-0.007(0.034)	-0.20	-0.017 (0.012)	-1.35
IRS	-0.038(0.032)	-1.17	0.035 (0.012)	2.97^{**}
FRS	0.084(0.021)	4.02^{***}	0.034 (0.008)	4.35^{***}
PMS	0.038(0.036)	1.06	-0.005 (0.016)	-0.30
Sex	0.103(0.057)	1.80	-0.127 (0.024)	-5.34^{***}
Age	0.033(0.013)	2.56^{**}	0.115 (0.004)	26.09^{***}

* IAUF: individual’s alcohol use frequency; SFIS: score for friend influence on individual’s smoking; FLIS: score for individual’s listening to his/her family; IRS: individual relationship score; FRS: family relationship score; PMS: parental marriage status

In GEE estimations for the marginal models given in Table 3, while the effects of alcohol use (IAUF) and FRS variables (p<0.001) and age (p<0.01) on smoking were found to be significant, effects of other independent variables and intercept (β_0) were found to be insignificant. In MI-GEE estimations for the marginal models given in Table 3, while effects of intercept (β_0), alcohol use (IAUF), FRS, sex and age variables (p<0.001) and IRS variable (p<0.01) on smoking were found to be significant, effects of other independent variables and were found to be insignificant. When the empirical standard errors obtained for missing and full data in Table 3 are noted, it is estimated that the standard errors for MI-GEE are smaller.

In GEE, even if the intra-individual relationships between the repeated responses are determined incorrectly, they allow for derivation of a consistent estimation of β. The empirical estimator requires only accurate modeling of the average response. This feature is important for longitudinal studies, because longitudinal studies are focused on the variation in average responses. Standard errors derived in the case of incorrect determination of intra-individual relationships are inapplicable. Therefore, the empirical (sandwich, empirical) variance estimator is used to derive standard errors applicable for β. For the model-based estimator, consistent estimations can be derived if both the average model and the working correlation matrix are derived accurately. Thus, the derivation of the applicable standard errors is possible (Wall *et al.* 2005; Halekoh, 2008; Lipsitz and Fitzmaurice 2009). Another important point in the results of the study is that the dispersion parameter was found to be 1.716 in the missing observation dataset, and 1.000 in the complete dataset. Therefore, when there are missing observations in the dataset the dispersion parameter can be higher than in the case of a complete dataset, and it is possible to think that excessive dispersion problems may be encountered. In addition, this result reveals the advantageous aspect of using the MI method (DeSouza *et al.* 2009: Quintano *et al.* 2010).

Conclusion

In conclusion, many situations should be considered in evaluation of longitudinal response variability. For the form of answer variability distribution, missing measurements structure, rate and analysis results, it is important to determine models (like classic approaches, GEE or GLMM) that will produce optimal information and to determine the correlation structure between repeated data or longitudinal data.

Acknowledgement

This study is a part of PhD thesis titled “Model Selection and Comparing Optimization Techniques in Marginal and Non-Marginal Multilevel Generalized Linear Mixed Model Using Missing Observed Longitudinal Data” of the first author.The thesis was funded by the project 2009-FBE-D079 of Scientific Research Projects Council of Van Yuzuncu Yil University. Additionally, an abstract of this paper was published as a book of abstracts in The 8th Conference of Eastern Mediterranean Region of the International Biometric Society (EMR 2015), May 11-15 2015 (pp: 125), in Cappadocia, Nevşehir, Turkey.

References

Aktaş, A., 2005. Genelleştirilmiş Eşitlik Kestirimi “GEE” (yüksek lisans tezi). Hacettepe Üniversitesi, Sağlık Bilimleri Enstitüsü, Ankara.

Allison, P. D., 2000. Multiple imputation for missing data: a cautionary tale. Sociological Methods and Research. 28:301–309.

Antonio, K., Beirlant, J., 2007. Actuarial statistics with generalized linear mixed models. Mathematics and Economics. 40: 58-76.

Barnett, A.G., Koper, N., Dobson, A. J., Schmiegelow, F., Manseau, M., 2010. Using information criteria to select the correct variance-covariance structure for longitudinal data in ecology. Methods in Ecology and Evolution. 1: 15-24.

Baygül, A., 2007. Kayıp Veri Analizinde Sıklıkla Kullanılan Etkin Yöntemlerin Değerlendirilmesi (yüksek lisans tezi). İstanbul Üniversitesi, Sağlık Bilimleri Enstitüsü, İstanbul.

Buuren, S., 2007. Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Reserach. 16: 219-242.

DeSouza, M.C., Legedza, T.R.A., Sankoh, A. J., 2009. An overview of practical approaches for handling missing data in clinical trials.

Journal of Biopharmaceutical Statistics. 19: 1055-1073.

Fitzmaurice, G. M., N.M. Laird, J. H. Ware, 2004. Applied Longitudinal Analysis. John Wiley and Sons Inc. 506p. New Jersey, USA.

Fitzmaurice, G., G. Verbeke, 2009. Parametric Modeling of Longitudinal Data: Introduction and Overview, Chap. 2. Longitudinal Data Analysis. Taylor and Francis Group, 618p. New York, USA.

Halekoh, U., 2008. Generalized estimating equations (GEE) lecture. http://gbi.agrsci.dk/statistics/courses/Rcourse-DJF2008/. (Erişim tarihi: 12.10.2010).

Hardin, J. W., J. M. Hilbe, 2003. Generalized Estimating Equations. Chapman and Hall/CRC. 218p.USA.

Hin, L. Y., Wang, Y. G., 2009. Working correlation structure identification in generalized estimating equations. Statistical in Medicine. 28: 642-658.

Jiang, J., 2007. Linear and Genaralized Linear Mixed Models and Their Applications. Springer Sciences+Business Media Inc. 251p. New York. USA.

Johnston, G., 1996. Repeated measures analysis with discrete data using the SAS system. http://www2.sas.com/proceedings/sugi22/STATS/PAPER278.PDF (Erişim Tarihi: 01.06.2010)

Lee, Y., Nelder, J. A., 2004. Conditional and marginal models: another view. Statist. Sci. 19;2 : 219–238.

Li, F., Maddalozzo, G. F., Harmer, P., Duncan, T.E., 1998. Analysis of longitudinal data of repeated observations using generalized estimating equations methodology. Measurement in Physical Education and Exercise Sciences. 2;2: 93-113.

Liang, K.Y., Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrika. 73;1: 13-22.

Lipsitz, S., G. Fitzmaurice, 2009. Generalized Estimating Equations for Longitudinal Data Analysis, Chap.3. Longitudinal Data Analysis. Taylor and Francis Group. 618p. New York, USA.

O’Brien, L.M., Fitzmaurice, G. M., 2005. Regression models for the analysis of longitudinal gaussian data from multiple sources. Stat. Med. 24;11: 1725-1744.

Park, T., Davis, C.S., Li, N., 1998. Alternative gee estimation procedures for discrete longitudinal data. Computational Statistics & Data Analysis. 28: 243-256.

Pekar, S., Brabec, M., 2018. Generalized estimating equations: A pragmatic and flexible approach to the marginal GLM modelling of correlated data in the behavioural sciences. Ethology. 124; 2: 86-93.

Quintano, C., Castellano, R., Rocca, A., 2010. Influence of outliers on some multiple imputation methods. Metodološki Zvezki. 7;1: 1-16

SAS, 2010. SAS/STAT Software: Hangen and Enhanced, Version 9.2, SAS, Inst. Inc., Cary, N.C. USA.

Schaffer, J. L., 1999. Multiple imputation: a primer. Statistical Methods in Medical Research. 8: 3–15.

Singer, J.D., J.B. Willet, 2003. Applied Longitudinal Data Analysis: Modeling Chance and Event Occurence. Oxford University Press, Inc. 627p. New York, USA.

Tan, T. K., Kang, T., Hogan, D., 2009. Using GEE to model student’s satisfaction: a SAS macro approach. Sas Global Forum, Statistics and Data Analysis. 251.

Wall, M.M., Dai, Y., Eberly, L. E., 2005. GEE estimation of a mis-specified time-varying covariate in Poisson regression with many observations. Statistics in Medicine. 24:925-939.

Yeşilova, A., Yılmaz, A., Kaki, B., 2006. Norduz erkek kuzularının bazı kesikli üreme davranış özelliklerinin analizinde doğrusal olmayan regresyon modellerin kullanılması. Yyü Tar Bil Derg 16; 2: 87-92.

Zeger, S. L., Liang, K.Y., Albert, P. S., 1988. Models for longitudinal data:a generalized estimating equation approach. Biometrics. 44;4: 1049-1060.