GU J Sci 31(1): 273-280 (2018)

Gazi University

Journal of Science

http://dergipark.gov.tr/gujs

Generalized Estimating Equations for Genetic Association Studies of Multi-Correlated Longitudinal Family Data

Özge KARADAĞ[1,*], Serpil AKTAŞ[1]

[1]*Department of Statistics, Hacettepe University, Ankara, Turkey*

Article Info

Received: 09/05/2017
Accepted: 20/11/2017

Keywords

genetic association
pedigree data
multi-correlated data
generalized estimating equations

Abstract

In genetic epidemiology studies, many diseases are multifactorial that can be both environmental and genetic inherited pattern. The relationship between genetic variability and individual phenotypes is usually investigated by genetic association studies. In genetic association studies, longitudinal measures are very important scale in detecting disease variants. They enable to observe both factors in the progress of disease. Generalized Linear Modelling (GLM) techniques offer a flexible approach for testing and quantifying genetic associations considering different types of phenotype distributions. In this study, it is aimed to accommodate Generalized Estimating Equations (GEE) method for genetic association studies in the presence of both familial and serial correlation. For this purpose, a real genotyped data set with the pedigree information and a continuous trait measured over time is used to model the association between the disease and the genotype by analyzing several variants, which have been associated with the disease. A joint working correlation structure is adapted, accounting for two different sources of correlations for estimating equations.

1. INTRODUCTION

In genetic epidemiology, many diseases are multifactorial because of the complex interaction of both genes and environment. Genetic association studies are used to analyze this complex relationship by testing the association between the disease and the genetic variation. In genetic association studies, complex designs are efficient and popular tools for providing substantive information however accompanying correlated structures. For instance, longitudinal measures, which induce a serial correlation, are important tools for detecting disease variants. In addition to that family study designs, which cause a familial correlation, are another necessary instrument for finding out the association between trait and the inherited genetic markers among the pedigree members. From this point of view, for testing the genetic association, a longitudinal pedigree framework will be useful and the data structure comes to a state of a multi-clustered data.

In the presence of clustered data, due to population stratification, the assumption on independent observations, which is required for maximum likelihood estimation, is often violated. Advanced models such as mixed and multilevel models are used for revealing the possible similarities among the subjects. GEE method, which is an extension to GLM, is often employed to analyze longitudinal and other correlated response data.

GEE is used to model clustered data especially occur in longitudinal structures where the measurements are collected from same individuals repeatedly. In such a case the source of the correlation is single. However in most of the genetic epidemiology studies, data reserve more than one source of correlation.

A method based on GEE for analysis of outcomes with multiple source of correlations was presented by Shults et al. [1]. The authors implemented the quasi-least squares method for analyzing the physical

*Corresponding author, e-mail: ozgekaradag@hacettepe.edu.tr

activity data which reserves three level of correlation. The presented method allows for some specific correlation structures, which are inadequate for genetic relatedness correlations. Especially in pedigree-based association analysis, the correlation structure should be user-defined rather than a specific structure such as exchangeable, autoregressive, etc.

In this study, we aim to accommodate GEE method for genetic association studies in the presence of both familial and serial correlation by defining a joint working correlation structure including both sources of correlation. For the serial correlation the known covariance structures are taken into consideration however for the familial correlation, the real kinship matrices are considered rather than a specific structure.

In the following section a brief introduction and some notations about GEE will be given and then the joint working correlation structure in the presence of multi-correlated data will be described. Next the association model for detecting the relationship between the genotype and the phenotype will be presented by implementing the joint working correlation for the longitudinal family framework. Following the theoretical part, a real data from GAW project will be analyzed by calculating the joint working correlation matrix which is consisting of the exact genetic relationship correlations in addition to serial correlations within the measurements. Finally, the analysis results will be discussed by making some concluding remarks.

2. GENERALIZED ESTIMATING EQUATIONS

GEE method was first introduced by Liang and Zeger [2] as a tool for analyzing covariance patterns. Covariance pattern models allow to define the type of the correlation between the responses of a given subject like linear mixed models. However, GEE method is more preferable in some cases compare to linear mixed modeling approach because of not requiring distributional assumptions. Moreover, it has a less complicated computation process. Especially when multiple correlated structures in which all levels of the correlation should be considered occur, GEE method provides an easier and efficient solution. Since the parameter estimation results of GEE are stable even if the correlation structure is mis-specified.

Actually, GEE itself is not a modeling technique, is an inference method for clustered data which treats the covariance as a nuisance factor and focuses on the mean model of the dependent variable. The models which based on the GEE method are called as population average or marginal models. As Hubbard et al. [3] pointed out, these kind of models describe the variability in the population mean given changes in covariates while accounting for the correlation among the dependent observations.

The within-subject correlation is incorporated into the model by obtaining a working correlation matrix. The working correlation matrix can be assigned as any of correlation structures. These are mostly independence, compound symmetry, unstructured and first order autoregressive structures.

Let $\boldsymbol{y}_i = (y_{i1,} y_{i2,} \dots, y_{iT})'$ be the $T \times 1$ response vector and similarly $\boldsymbol{X}_i = (\boldsymbol{X}_{i1,} \boldsymbol{X}_{i2,} \dots, \boldsymbol{X}_{iT})'$ be the $T \times p$ design matrix where $\boldsymbol{X}_{it}$ is covariate vector of the i^{th} subject measured at time t, $(i = 1,2, \dots, N, t = 1,2, \dots, T)$. Following these notations, the mean model of the i^{th} subject, $E(\boldsymbol{y_i}|\boldsymbol{X}_i) = \boldsymbol{\mu}_i$ can be written as using the marginal generalized linear model equation

$$g(E(\boldsymbol{y}_i)) = g(\boldsymbol{\mu}_i) = \boldsymbol{X}_i\boldsymbol{\beta} \tag{1}$$

where $g(.)$ is a known function, referred as the link function which transforms the mean $\boldsymbol{\mu}_i$ to the linear predictor in the presence of measurements from a distribution with density from exponential family (Gaussian, binomial, Poisson, gamma, etc.). $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_{p-1})'$ is the $p \times 1$ vector of fixed effect parameters and $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \dots, \mu_{iT})$ is the mean vector.

GEE method is based on alternating the updated estimates of fixed effect parameters and the correlation parameter by using an iterative procedure. In the first step, fixed effect parameters are estimated by using

the quasi likelihood while holding the variance covariance parameters constant. In the second step, the variance covariance parameters are estimated while holding fixed ones constant and in the next step these parameters are fixed to re-estimate the fixed parameters. These last two steps are repeated until the parameter estimates convergence [4].

Variance covariance matrix of $\boldsymbol{y}_i$, $\boldsymbol{V}_i$, that is specified through the working correlation matrix $R_i(\alpha)$ is estimated by

$$\widehat{\boldsymbol{V}}_i = \phi \boldsymbol{A_i}^{\mathbf{1/2}} R_i(\alpha) \boldsymbol{A_i}^{1/2} \tag{2}$$

where $\boldsymbol{A_i} = \mathrm{diag}(\mathrm{h}(\mu_{\mathrm{i1}}), \ldots, \mathrm{h}(\mu_{\mathrm{iT}}))$ is the diagonal matrix of known variance functions $\mathrm{h}(.)$ and ϕ is the dispersion/scale parameter.

The general framework of GEE is defined as

$$\sum_{\boldsymbol{i}=1}^{N} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)' \widehat{\boldsymbol{V}}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu_i}) = 0 \tag{3}$$

GEE method is generally generated to model clustered data when the source of correlation is single. However in genetic applications usually multi-correlated data occurs due to the stratified pedigree structure. In the following section, the methodology of GEE is accommodated for multi-correlated data structures.

3. GEE FOR MULTI-CORRELATED DATA

In the presence of multiple correlated structured data, the clusters should be defined carefully. For instance, in a longitudinal structure, the cluster determiner is the individual whereas the cluster members are repeated measurements. On the other hand, in a two level correlated structure such as longitudinal pedigree structure, this time the cluster identifier will be pedigrees and the clusters arise from the repeated measurements of the same pedigree members.

Random effects or mixed effects linear models are used as a solution in the presence of multi-level correlated data [5,6]. Latent variable modelling is another approach dealing with the multiple source of correlation [7]. In recent years, GEE algorithm is also used as a solution in the presence of multiple source of correlation. The GEE method was first applied to account for multi-level correlation by Shults et al. [1] implementing quasi least squares with a generalization of correlation structure proposed by Lefkopoulou [8]. Compare to alternative methods such as random effects modelling, GEE approach has advantages because of not requiring distributional assumptions.

The GEE algorithm for analysis of multiple correlated responses with two or more sources of correlation is basically based on a joint working correlation structure which is obtained by the Kronecker product of all potential correlation structures.

3.1. Longitudinal Family Data

In a longitudinal pedigree framework, there are two source of correlation. Supposing the working correlation matrix for the serial correlation within each subject is represented by $\mathrm{R_s}(\alpha)$ and within each pedigree familial correlation matrix is represented by $\mathrm{R_f}(\alpha)$. Then the joint correlation matrix structure can be written as $\mathrm{R}(\alpha) = \mathrm{R_s}(\alpha) \otimes \mathrm{R_f}(\alpha)$.

In genetic association studies, the familial correlation within each pedigree is defined by the genetic relatedness matrix, $\mathbf{G}$. Genetic relatedness matrix is a symmetric block diagonal matrix consists of the kinship coefficients. The pairwise genetic relatedness matrix $\mathbf{G}$ is consist of genetic correlation

coefficients between any two individuals and calculated as $\mathbf{G} = 2\boldsymbol{\varphi}$ where $\boldsymbol{\varphi}$ is the kinship matrix. In family-based studies, the size of the pedigrees usually differs from each other. In a sample consisting of m pedigrees, let j^{th} pedigree has n_j individuals and φ_{ik} individually reflects the pairwise kinship coefficient between individuals i and k in the j^{th} pedigree, $(j = 1,2 \dots, m)$ and $(i, k = 1,2 \dots, n_j)$. The kinship coefficient φ_{ik} is the measure of the probability of two alleles, one sampled at random from each individual are identical by descent. Supposing the partial working correlation matrix for the j^{th} pedigree is $R_{f(j)}(\alpha) = \mathbf{G}_j$, then the joint working correlation for both fixed serial and pedigree based familial relatedness, can be written as

$$R(\alpha) = \begin{bmatrix} R_s(\alpha)\otimes\mathbf{G}_1 & 0 & 0 & \dots & 0 \\ 0 & R_s(\alpha)\otimes\mathbf{G}_2 & \vdots & \dots & 0 \\ 0 & 0 & R_s(\alpha)\otimes\mathbf{G}_3 & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & R_s(\alpha)\otimes\mathbf{G}_m \end{bmatrix} \quad (4)$$

Thus, in a longitudinal pedigree framework the joint working correlation is a (Nt)x(Nt) symmetric block diagonal matrix, where t is the number of replications, n_j is size of the j^{th} pedigree and $N = \sum_{j=1}^{m} n_j$ is total number of individuals.

3.2. Genetic Association Model

In this paper, we use GEE method as a solution for testing the association between candidate genes and phenotype by using a longitudinal pedigree framework in which two source of correlation should be considered, one arises from the repeated measurements and the other is derived from the familial structure. In genetic analysis, the individual genetic information of a single variant, SNP (Single Nucleotide Polymorphism), is integrated into association model by the genotype measurement. SNP is defined as the variation in a single nucleotide at a specific position on genome that causes the individual diversity and individual genotype correspondences to number of minor alleles in the allelic combination of relevant SNP. For analyzing the relationship between the SNP and the longitudinal blood pressure phenotype, a SNP based generalized linear mixed model is obtained as follows

$$y_{itj(s)} = \boldsymbol{X}_{itj}\boldsymbol{\beta} + snp_s + u_j + g_{ij} + w_{it} + e_{itj} \quad (5)$$

where $i = 1, \dots, N$, $t = 1, \dots, T$, $j = 1, \dots, m$ and $s = 1, \dots, S$. y_{itj} is the response of t^{th} repeated measure of i^{th} individual in the j^{th} pedigree and $\boldsymbol{X}_{ijt}$ is the covariate design matrix corresponding to $\boldsymbol{\beta}$ fixed effect coefficients parameter vector. snp_s represents the fixed effect of the s^{th} SNP. u_j, g_{ij} and w_{it} are the random effects which are representing the common environmental factor for the j^{th} pedigree, genetic correlation among individuals of the j^{th} pedigree and serial correlation among repeated measurements of the i^{th} individual, respectively. e_{itj} is the error term.

The joint working correlation matrix, related to association model given by (5) can be given as

$$R_s(\alpha)\otimes\mathbf{G}_j = R_s(\alpha)\otimes\begin{bmatrix} 1 & G_{j(12)} & G_{j(13)} & \dots & G_{j(1m_j)} \\ & 1 & G_{j(23)} & \dots & G_{j(2m_j)} \\ & & 1 & \dots & \vdots \\ & & & \ddots & \vdots \\ & & & & 1 \end{bmatrix} \quad (6)$$

where $G_{j(k,l)}$ is the genetic correlation coefficient between the members k and l of pedigree j, including m_j individuals $(k, l = 1,2, \dots, m_j)$. The number of elements in the upper-triangle of the joint working correlation matrix can be calculated as

$$\sum_{j=1}^{m} m_j t \frac{(m_j t - 1)}{2} \tag{7}$$

where *t* is the number of replicates.

4. ANALYSIS TO GAW PROJECT DATA

GAW (Genetic Analysis Workshop) is a collaborative effort among genetic epidemiologist to evaluate and compare statistical genetic methods. Real human genome sequence and longitudinal phenotype data are provided for 20 large Hispanic pedigrees/families including different number of individuals, the smallest pedigree includes 27 individuals where as the largest has 107 with a total of 1389. In GAW project main interest is to find candidate genes or genome regions that contribute to blood pressure, which is a hereditary disease. Systolic blood pressure (SBP), measured while the heart is beating, has been identified as a risk factor for cardiovascular events such as kidney failure and heart attack. For high blood pressure a genetic component has been found out by several genome-wide association studies such as Levy et al. [9].

The International Consortium for Blood Pressure Genome-Wide Association Studies used a multi-stage design in 200000 individuals of European pedigrees and published 16 loci that have been associated with hypertension (SBP≥140 mm Hg). In this paper we focus on bi allelic variants in 3 novel genes *MOV10, ADM* and *CYP1A1* which have been reported as associated genes with SBP. Gene *MOV10* is located in chromosome 1 between position 112674312 and 112700746. Gene *ADM* is located in chromosome 11 between position 10304980 and 10307402 and gene *CYP1A1* is located in chromosome 15 between position 74719542 and 74725610. GAW data was released for odd number chromosomes and includes 180, 14 and 36 genotyped variants respectively, for each gene above-mentioned.

In projects that GAW data drawn from, participants are followed in a longitudinal design. Through the study, participants took successive examinations. The phenotype of interest SBP and the time-dependent covariates age, anti-hypertensive medication use and smoking statuses were obtained for three successive visits by using the real pedigrees and the cleaned imputed sequence data and were based on the real blood pressure distributions. The elapsed times between three visits are 3.9 and 3 years, respectively. Longitudinal phenotype and the genotype data are available for 849 individuals [10].

5. RESULTS

Under 3 novel genes ADM, CYP1A1 and MOV10, 230 SNPS have been analyzed by the adapted GEE method by adjusting for the medication, age and smoking status effects. Due to the large amount of analyzed SNPs, it is preferred to summarize the logarithmic transformed p-values of each SNP association model by using a Manhattan plot as given in Figure 1.

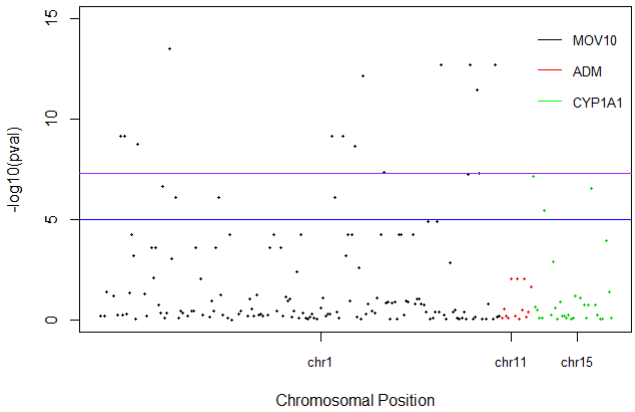

Figure 1. *Manhattan Plot*

Table 1. *Parameter estimates for featured genome-wide significant SNPs by using GEE*

			Intercept		age		medication		smoke		SNP	
RefSNP	Chr	Position	**Coef.**	**P>|W|**	**Coef.**	**P>|W|**	**Coef.**	**P>|W|**	**Coef.**	**P>|W|**	**Coef.**	**P>|W|**
rs6668939	1	112675387	109,3494	<0.001*	0,3277	<0.001*	13,8409	<0.001*	0,0744	0,9330	2,4934	7,04e-10*
rs6668952	1	112675421	109,3494	<0.001*	0,3277	<0.001*	13,8409	<0.001*	0,0744	0,9330	2,4934	7,04e-10*
rs148615626	1	112675912	109.3327	<0.001*	0,3285	<0.001*	13,8370	<0.001*	0,1056	0,9100	-3,3110	1,8e-09*
rs80330026	1	112677902	109,3320	<0.001*	0,3280	<0.001*	13,8626	<0.001*	0,0592	0,9500	4,7658	3,2e-14*
rs74109943	1	112688126	109,3494	<0.001*	0,3277	<0.001*	13,8409	<0.001*	0,0744	0,9330	2,4934	7,04e-10*
rs74109945	1	112688567	109,3494	<0.001*	0,3277	<0.001*	13,8409	<0.001*	0,0744	0,9330	2,4934	7,04e-10*
rs147967092	1	112689340	109,3585	<0.001*	0,3276	<0.001*	13,8545	<0.001*	0,0916	0,9200	-2,2181	2,1e-09*
not identified	1	112689715	109,3068	<0.001*	0,3285	<0.001*	13,8242	<0.001*	0,0385	0,9700	5,2992	6,6e-13*
rs562477917	1	112690410	109,3912	<0.001*	0,3269	<0.001*	13,8859	<0.001*	0,0868	0,9200	-1,8151	4,7e-08*
rs145675628	1	112693068	109,3353	<0.001*	0,3280	<0.001*	13,8751	<0.001*	0,0765	0,9300	1,9911	1,9e-13*
rs190432897	1	112694904	109,3353	<0.001*	0,3280	<0.001*	13,8751	<0.001*	0,0765	0,9300	1,9911	1,9e-13*
not identified	1	112695538	109,3665	<0.001*	0,3272	<0.001*	13,8632	<0.001*	0,0876	0,9200	1,9553	3,5e-12*
rs534481311	1	112696740	109,3353	<0.001*	0,3280	<0.001*	13,8751	<0.001*	0,0765	0,9300	1,9911	1,9e-13*
rs528140549	15	74722721	109,3341	<0.001*	0,3282	<0.001*	13,7905	<0.001*	0,0910	0,9200	4,2085	1,8e-10*

* has significant effect on SBP at 0.05

In Figure 1, the straight purple line corresponds to required genome-wide significance threshold value, $p < 5 \times 10^{-8}$ and blue line corresponds to suggestive significance threshold value, $p < 1 \times 10^{-5}$ [11]. There are 23 SNPs detected above the suggestive threshold value and 14 SNPS above the genome-wide threshold value out of 230. Within the genome-wide significant SNPs 13 of them are from MOV10 gene which is located in chromosome 1 and one of them is from CYP1A1 gene from chromosome 15. For ADM gene none of the variants are found to be associated with SBP in genome-wide significance level.

For the detailed discussion of the model results we focused on 14 genome-wide associated SNPs from two novel genes, MOV10 and CYP1A1. Most of the variants have been identified by dbpSNP database due to the Genome Reference Consortium GRCh37. Only two variants out of 14 hasn't been identified by any databases. These SNPs are from chromosome 1 with base pair positions 112689715 and 112695538. The parameter estimation results and available reference SNP numbers of featured variants are given in Table 1.

Based on the GEE results, medication and age are found to have significant effects on SBP, otherwise smoking status has no substantive influence. From Table 1, it can be drawn as a conclusion that for some variants, the parameter estimation results are close to each other due to the having same genotype distribution.

6. CONCLUSION

In this study we analysed 230 variants from three novel genes, which have been already reported to be associated with SBP, by accounting both familial and serial correlations between observations. A GEE algorithm was adapted and applied to multi-correlated genetic data structure which allows for a user-defined joint working correlation matrix. Genetic association of each SNP was tested by using adapted algorithm. The effects of covariates, medication, age and smoking status were also considered. Comparing to other GEE methods, the adapted GEE method allows for accounting the real genetic relatedness coefficients in the presence of repeated phenotype measurements.

CONFLICTS OF INTEREST

No conflict of interest was declared by the authors.

REFERENCES

[1] Shults, J., Whitt, M.C., Kumanyika, S., "Analysis of data with multiple sources of correlation in the framework of generalize estimating equations", Statistics in Medicine, 23(20): 3209- 3226, (2004).

[2] Liang, K.Y., Zeger, S.L., "Longitudinal data analysis using generalized linear models", Biometrika, 73: 13-22, (1986).

[3] Hubbard, A.E., Ahern, J., Fleischer, N.L., Van der Laan, M., Lippman, S.A., Jewell, N., Bruckner, T., Satariano, W.A., "To GEE or Not to GEE: Comparing population average and mixed models for estimating the associations between neighbourhood risk factors and health", Epidemiology, 21(4): 467- 474, (2010).

[4] Brown, H., Prescott, R., Applied Mixed Models in Medicine. Statistics in Practice, Wiley Series, England, (2006).

[5] Qaqish, B. F. and Liang K. Y., "Marginal models for correlated binary responses with multiple classes and multiple levels of nesting", Biometrics, 48, 939-950,(1992).

[6] Ten Have, T.R., Kunselman, A.R., Tran, L., "A comparison of mixed effects logistic regression models for binary response data with two nested levels of clustering", Statistics in Medicine, 18(8): 947-960, (1999).

[7] Joy, J. and Lin X., "Latent Variable Models for Longitudinal Data with Multiple Continuous Outcomes", Biometrics, 56(4):1047-1054, (2000).

[8] Lefkopoulou, M. Moore, D., Ryan, L. " The Analysis of Multiple Correlated Binary Outcomes: Application to Rodent Teratolgy Experiments", Journal of the American Statistical Association, 84(407): 810-815, (1989).

[9] Levy, D., Ehret, G.B., Rice, K., Verwoert, G.C., Launer, L.J., Dehghan, A., Glazer, N.L., Morrison, A.C., Johnson, A.D., Aspelund, T., Aulchenko, Y., Lumley, T., Köttgen, A., Vasan, R.S., Rivadeneira, F., Eiriksdottir, G., Guo, X., Arking, D.E., Mitchelli,.G.F., Mattace-Raso, F.U., Smith, A.V., Taylor, K., Scharpf, R.B., Hwang, S.J., Sijbrands, E.J., Bis, J., Harris, T.B., Ganesh, S.K., O'Donnell, C.J., Hofman, A., Rotter, J.I., Coresh, J., Benjamin, E.J., Uitterlinden, A.G., Heiss, G., Fox, C.S., Witteman, J.C., Boerwinkle, E., Wang, T.J., Gudnason, V., Larson, M.G., Chakravarti, A., Psaty, B.M., van Duijn, C.M., "Genome-wide association study of blood pressure and hypertension", Nat. Genet., 41(6): 77–687, (2009).

[10] Almasy, L., Dyer, T.D., Peralta, J.M., Jun, G., Wood, A.R., Fuchsberger, C., Almeida, M.A., Kent, Jr S.W., Fowler, S., Blackwell, T.W., Puppala, S., Kumar, S., Curran, J.E., Lehman, D., Abecasis, G., Duggirala, R., Blangero, J., The T2D-GENES Consortium, "Data for genetic analysis workshop 18: human whole genome sequence, blood pressure and simulated phenotypes in extended pedigrees", BMC Proc., 8 (suppl. 2):S2, (2014).

[11] Barsh, G.S., Copenhaver, G.P., Gibson, G., Williams, S.M., "Guidelines for genome-wide association studies", PloS Genetics, 8, 7, (2012).