Estimation of growth curve models with structured error covariances by generalized estimating equations

Heungsun Hwang

HEC Montreal

Yoshio Takane

McGill University

The work reported in this paper was supported by Grant 290439 and Grant A6394 from the Natural Sciences and Engineering Research Council of Canada to the first and second authors, respectively. We thank Terry Duncan for his alcohol use data. The alcohol use data were made available by Grant DA09548 from the National Institute on Drug Abuse. Requests for reprints should be sent to: Heungsun Hwang, Department of Marketing, HEC Montreal, 3000 Chemin de la Côte Ste Catherine, Montreal, QC, H3T-2A7, Canada. Email: heungsun.hwang@hec.ca.

Abstract

The growth curve model is useful for the analysis of longitudinal data. It helps investigate an overall pattern of change in repeated measurements over time and the effects of time-invariant explanatory variables on the temporal pattern. The traditional growth curve model assumes that the matrix of covariances between repeated measurements is unconstrained. This unconstrained covariance matrix often appears unattractive. In this paper, the generalized estimating equation method is adopted to estimate parameters of the growth curve model. As a result, the proposed method allows a more variety of constrained covariance structures than the traditional growth curve model. An empirical application is provided so as to illustrate the proposed method.

Key Words: longitudinal data, quasi-likelihood, QIC, alcohol use.

1. Introduction

The growth curve model (Grizzle & Allen, 1969; Khatri, 1966; Potthoff & Roy, 1964; Rao, 1965) has been used for the analysis of longitudinal data, in which measurements are repeatedly taken on a response variable at a number of time points. This method enables us to investigate an overall pattern of change in the response variable over time. It also allows us to examine the effects of time-invariant explanatory variables on the temporal pattern of the response variable. The traditional growth curve model assumes that the covariance matrix of repeated measurements is unconstrained or unstructured. However, this unconstrained covariance matrix is computationally less attractive when the number of time points becomes large or when the number of repeated measurements is not the same across individuals, often leading to less reliable parameter estimates (Laird & Ware, 1982; Duncan, Duncan, Hops, & Stoolmiller, 1995).

Instead, the covariance matrix may be assumed to be constrained in a certain way. If the constraints imposed on a covariance matrix are consistent with the data, one may obtain more reliable estimates of parameters, reducing the number of the covariances to be estimated. Moreover, one may obtain empirically more insightful information on the relationship between repeated measurements through such a constrained covariance matrix. For instance, according to Bagozzi and Bergami (2002), workers tend to identify themselves most strongly with their coworkers, somewhat less strongly with the job or task, still less strongly with their department, and least strongly with the organization. A first-order autoregressive structure (Jöreskog, 1970) or a simplex pattern (Guttman, 1954) within the covariances among the different targets of identification may be assumed to

test the hypothesis of such a linear dependence ordered in a sequence (Bagozzi & Bergami, 2002).

In practice, however, the covariance structure of repeated measurements is usually unknown, so that it is difficult to decide which constrained structure is tenable in advance. This may often lead to the misspecification of the covariance matrix structure, thereby distorting the results of parameter estimation (Diggle, 1998). The generalized estimating equation (GEE) method (Liang & Zeger, 1986; Zeger & Liang, 1986) may be used for valid parameter estimation in such a situation. The GEE is an extension of the quasi-likelihood method (Wedderburn, 1974; McCullagh, 1983), which offers asymptotically consistent parameter estimates even if the covariance structure of repeated measurements is not correctly specified. It thus allows more diverse constrained structures in the population covariance matrix compared to standard linear models.

In this paper, it is proposed to use GEE for estimating the parameters of the growth curve model. As a result, the proposed method enables to accommodate various types of constrained covariance structures, which is advantageous over the traditional growth curve model. The paper is organized as follows. In Section 2, the proposed method is discussed in detail. In Section 3, an actual longitudinal data set is analyzed to demonstrate the usefulness of the proposed method. The final section is devoted to discussing several additional aspects of the proposed method.

2. The Proposed Method

Let $\mathbf{y}_i$ denote a T by 1 vector of measurements of individual i ($i = 1, \cdots, N$) on a single response variable across T time points. Let $\mathbf{x}_i$ denote a P by 1 vector of time-

invariant explanatory variables for individual *i*. The traditional growth curve model (Potthoff & Roy, 1964) is written as

$$\mathbf{y}_i = \mathbf{ABx}_i + \mathbf{e}_i, \tag{1}$$

where **B** is a *D* by *P* matrix of unknown coefficients, **A** is a *T* by *D* matrix of known basis functions that represent specific aspects of change in the response variable over the time points, and $\mathbf{e}_i$ is a *T* by 1 vector of measurement errors for individual *i*. It is assumed that $\mathbf{e}_i \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma}_i)$. In (1), the population covariance matrix $\mathbf{\Sigma}_i$ is assumed to be unconstrained, so that it contains *T(T-1)/2* distinct covariances and *T* variances among repeated measurements. This unconstrained covariance structure is not recommended in some cases (Laird & Ware, 1982; Duncan et al., 1995).

In this paper, the GEE method is utilized to estimate model parameters in (1) under more diverse covariance structures than the traditional growth curve model. The GEE is a quasi-likelihood method (Wedderburn, 1974; McCullagh, 1983), where the covariance matrix of $\mathbf{y}_i$ is specified by a working covariance matrix. Let $\mathbf{V}_i$ denote the working covariance matrix for $\mathbf{y}_i$, defined by

$$\mathbf{V}_i = \phi \mathbf{R}_i(\boldsymbol{\theta}), \tag{2}$$

where ϕ is a scale parameter, and $\mathbf{R}_i(\boldsymbol{\theta})$ is a working correlation matrix for $\mathbf{y}_i$. It is assumed that $\mathbf{R}_i(\boldsymbol{\theta})$ is a function of a *q* by 1 vector of $\boldsymbol{\theta}$, which is the same across all individuals. The $\mathbf{R}_i(\boldsymbol{\theta})$ can take various forms, depending on what kind of covariance structures are assumed in (1). The covariance structures commonly assumed in GEE are 1) independence, 2) exchangeable, 3) auto-regressive, and 4) unconstrained. The independence covariance structure represents no correlations among repeated

measurements. It is obtained simply when $\mathbf{R}_i(\boldsymbol{\theta}) = \mathbf{I}$. The exchangeable covariance structure indicates that all covariances between repeated measurements are identical, that is, $corr(y_{it}, y_{it'}) = \theta$ for all $t \neq t'$ ($t = 1, \cdots, T$). This kind of structure is also obtained from a random-effects model (e.g., Laird & Ware, 1982). If time points are unequal across individuals, the exchangeable covariance structure may be appropriate (Duncan et al., 1995). The auto-regressive covariance structure represents a correlation as a function of the time between two repeated measurements. The first-order auto-regressive (AR-1) structure is typically used for time-series data (Liang & Zeger, 1982; Duncan et al., 1995), which is equivalent to $corr(y_{it}, y_{it'}) = \theta^{|t-t'|}$ for all $t \neq t'$. The unconstrained structure leads to *T(T-1)/2* distinct correlations among repeated measurements. The unconstrained covariance structure appears suitable when the number of repeated measurements is small and is equal across individuals (Liang & Zeger, 1986; Duncan et al., 1995).

The GEE estimation procedure for (1) consists of two steps. In the first step, for fixed $\mathbf{V}_i$, the quasi-likelihood estimate of **B**, denoted by $\hat{\mathbf{B}}$, is obtained. Let **b** = vec(**B**), where vec(**U**) is a vector formed by stacking all columns of **U** one below another. Let $\hat{\mathbf{b}}$ denote the quasi-likelihood estimate of **b**. We note that $\mathbf{y}_i - \mathbf{A}\mathbf{B}\mathbf{x}_i = \text{vec}(\mathbf{y}_i - \mathbf{A}\mathbf{B}\mathbf{x}_i) = \mathbf{y}_i - (\mathbf{x}_i' \otimes \mathbf{A})\mathbf{b} = \mathbf{y}_i - \boldsymbol{\Omega}_i \mathbf{b}$, where $\boldsymbol{\Omega}_i = \mathbf{x}_i' \otimes \mathbf{A}$. Then, $\hat{\mathbf{b}}$ is the solution of a quasi-score function

$$\sum_{i=1}^{N} \boldsymbol{\Omega}_i' \mathbf{V}_i^{-1} (\mathbf{y}_i - \boldsymbol{\Omega}_i \mathbf{b}) = \mathbf{0}. \tag{3}$$

Thus, solving (3) for **b** yields

$$\hat{\mathbf{b}} = \left\{ \sum_{i=1}^{N} \boldsymbol{\Omega}_i' \mathbf{V}_i^{-1} \boldsymbol{\Omega}_i \right\}^{-1} \left\{ \sum_{i=1}^{N} \boldsymbol{\Omega}_i' \mathbf{V}_i^{-1} \mathbf{y}_i \right\}. \tag{4}$$

The updated $\mathbf{B}$ is reconstructed from $\hat{\mathbf{b}}$. In the next step, ϕ and $\mathbf{R}_i(\boldsymbol{\theta})$ are estimated by the current Pearson residuals (McCullagh & Nelder, 1989, p. 37), defined as follows:

$$\boldsymbol{\eta}_i = \mathbf{y}_i - \mathbf{A}\hat{\mathbf{B}}\mathbf{x}_i. \tag{5}$$

Then, ϕ is updated by

$$\hat{\phi} = \frac{1}{N-K}\sum_{i=1}^{N}\boldsymbol{\eta}_i'\boldsymbol{\eta}_i, \tag{6}$$

where K is the number of parameters in $\mathbf{B}$. In turn, $\boldsymbol{\eta}_i$ and ϕ are used to update $\mathbf{R}_i(\boldsymbol{\theta})$. For specific derivations of $\mathbf{R}_i(\boldsymbol{\theta})$ for the constrained covariances stated above from $\boldsymbol{\eta}_i$ and ϕ, refer to Liang and Zeger (1986).

These two steps are alternated until convergence is reached. Once $\mathbf{B}$ is estimated, its asymptotic covariance estimates can also be obtained. If the covariance structure is correctly specified, the consistent estimates of the asymptotic covariances of $\mathbf{B}$ is given by

$$\boldsymbol{\Gamma} = \left\{\sum_{i=1}^{N}\boldsymbol{\Omega}_i'\mathbf{V}_i^{-1}\boldsymbol{\Omega}_i\right\}^{-1}. \tag{7}$$

The values in (7) are called the naive covariance estimates. On the other hand, if the covariance structure is misspecified, the asymptotic covariances estimates of $\mathbf{B}$ is given by

$$\boldsymbol{\Psi} = \boldsymbol{\Gamma}\left\{\sum_{i=1}^{N}\boldsymbol{\Omega}_i'\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{A}\mathbf{B}\mathbf{x}_i)(\mathbf{y}_i - \mathbf{A}\mathbf{B}\mathbf{x}_i)'\mathbf{V}_i^{-1}\boldsymbol{\Omega}_i\right\}\boldsymbol{\Gamma}. \tag{8}$$

The values in (8) are called the robust covariance estimates because they are consistent even if the structure of $\boldsymbol{\Sigma}_i$ is misspecified. A well-known advantage of the GEE methodology is that the working correlation matrix need not be correctly specified to

obtain asymptotically consistent parameter estimates since it relies only on correct specification of the marginal expectation or mean model, treating covariances as nuisance parameters. If the working correlation matrix is correctly specified, however, the resultant parameter estimates are efficient. Moreover, GEE provides asymptotically consistent covariance estimates of the parameter estimates even if the covariance matrix is not correctly specified.

The proposed method allows fitting a broad range of models with different temporal patterns of change as well as different covariance structures. Due to the absence of the actual likelihood, however, it is difficult to apply likelihood-based information criteria such as AIC and BIC for model selection in GEE. Instead, one may use Pan (2001)'s QIC (the quasi-likelihood under the independence model criterion). QIC is a modification to AIC for GEE, where the likelihood function value in AIC is replaced by the quasi-likelihood function value obtained under $\mathbf{R}_i(\boldsymbol{\theta}) = \mathbf{I}$ and the penalty term is adjusted. QIC is defined as

$$\text{QIC} = -2Q(\mathbf{B}) + 2\text{trace}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Psi}), \tag{9}$$

where Q(**B**) is the value of the quasi-likelihood under the independence assumption, computed by the GEE estimator of **B** based on any working correlations. The second term in (9) reflects the degree of the differences between the naive and robust covariance estimates of **B**, which indicates how much the working covariance matrix is consistent with the true covariance matrix (i.e., the smaller the differences are, the more consistent the covariance matrices are with each other) (Zeger & Liang, 1986). For the normal case, specifically, the quasi-likelihood under the independence assumption is given by $-\sum_{i=1}^{N}(\mathbf{y}_i - \mathbf{ABx}_i)'(\mathbf{y}_i - \mathbf{ABx}_i)/2$ (McCullagh & Nelder, 1989, p. 326). Like AIC, a

model that minimizes QIC is regarded as the most appropriate one among fitted models. Pan (2001) showed that QIC performed well for model selection in GEE.

3. Empirical Application

The present example is part of a longitudinal survey on the predictors and consequences of substance use among adolescents from American northwestern urban areas (Duncan, Duncan, Alpert, Hops, Stoolmiller, & Muthén, 1997). The sample consists of 632 adolescents measured on their use of alcohol over 4 time points. The measure of alcohol use was assessed based on a self-reported 5-point item: (1) life time abstainers, (2) 6-month abstainers, (3) current use of less than four times a month, (4) current use of between 4 and 29 times a month, and (5) current use of 30 or more times a month. Five additional variables were measured once at the initial time point, including parental marital status, family status, socio-economic status (SES), age, and gender. Marital status was classified as follows: 0 = single and 1 = married or living in a committed relationship. Family status was categorized as follows: 0 = step or foster families and 1 = others. SES was calculated as the average of parental annual income and education level. Parental annual income was assessed based on a 16-point scale ranging from '6,000 dollars and below' to '50,000 dollars or more'. Education levels range from 'Grade level 6 or less' to 'Graduate level'. Male and female were coded as 0 and 1, respectively. The measure of alcohol use was used as $\mathbf{y}_i$. The five additional variables were considered as explanatory variables in $\mathbf{x}_i$. In addition, $\mathbf{x}_i$ contained an intercept term.

Table 1 shows descriptive statistics of the response variable measured at four time points.

Insert Table 1 about here

From the descriptive statistics it seems that there existed a linear trend of change in the response variable because its mean levels were likely to increase monotonically over time. In addition, the response variable was merely measured over four time points. This relatively small number of time points may not be sufficient to reveal a complex non-linear temporal pattern in the response variable (MacCallum, Kim, Malarkey, & Kiecolt-Glaser, 1997). According to these prior investigations about the data, we initially assumed a linear-trend model where **A** was defined as a known matrix of orthogonal polynomials of order 1 to represent a linear trend of temporal change in alcohol use (the exact form of the orthogonal polynomials are given below). We also considered alternative models with two different temporal patterns of change such as a quadratic trend and no time-specific trend or stability over time. Furthermore, we assumed four types of covariance structures for each of the models: Independence, exchangeable, AR-1, and unconstrained.

Table 2 provides a summary of the goodness of fit of the fitted models.

Insert Table 2 about here

In Table 2, model 1 corresponds with a model in which alcohol use was assumed to vary in a quadratic fashion over the time points. In model 1, **A** was pre-specified as a matrix of orthogonal polynomials of order 2, that is,

$$\mathbf{A} = \begin{bmatrix} 1 & -3 & 1 \\ 1 & -1 & -1 \\ 1 & 1 & -1 \\ 1 & 3 & 1 \end{bmatrix}. \tag{14}$$

Model 2 is the linear-trend model that we originally assumed. The last model (model 3) specified **A** as a matrix of orthogonal polynomials of zero order, assuming that there was stability in the response variable over time. Each of the models was associated with the four types of covariance structures. We chose the linear-trend model with the AR-1 covariance structure as the final model because it showed the smallest QIC value among the fitted models. The final model posits that the measurements on alcohol use are likely to vary in a linear manner during the study. This linear pattern of change in alcohol use of the same sample was also reported by Duncan et al. (1997). Furthermore, it shows that the covariances among the repeated measurements seem to be structured in a first-order auto-regressive or a simplex pattern. It implies that the repeated measure at time point *t* is affected by that at time point *t-1*.

Table 3 provides the GEE estimates of **B** obtained from the best fitting model, along with their robust standard errors in the parentheses.

Insert Table 3 about here

The first column of **B** under the label of *Initial* provides the effects of the explanatory variables on the response variable at the initial status, and the second column under the label of *Linear* represents the effects of the explanatory variables on a growth rate of the response variable over time. As shown in Table 3, the estimated intercepts for Initial and

Linear are all significant and positive. This indicates that there exist a significant level of alcohol use at the initial status and a significant linear growth rate of alcohol use over the four time points. Parental marital status has a significant and negative impact on the initial status of alcohol use, indicating a higher level of alcohol consumption by adolescents living with single parents at the initial assessment compared to those living with both parents. Family status also has a significant and positive effect on the initial status of alcohol use. This suggests that adolescents living with other families rather than step or foster families show higher levels of alcohol use at the initial assessment. On the other hand, age exhibits a significant and positive effect on the growth rate of alcohol use, indicating that older adolescents tend to increase alcohol consumption at a higher rate compared to younger adolescents over the four assessments. Finally, gender shows a significant and positive impact on the initial status and a significant and negative effect on the growth rate of alcohol use. This suggests that male adolescents appear to consume a higher level of alcohol at the initial status while female adolescents tend to show a higher growth rate of alcohol use over the four assessments.

4. Concluding Remarks

The GEE methodology was typically proposed for consistent estimation of the effects of explanatory variables on repeated measurements, taking into account the dependency among the repeated measurements. In this paper, the GEE was adopted for the parameter estimation of the growth curve model. As such, the proposed method enables to investigate an overall temporal pattern of change in a response variable and the effects of explanatory variables on the temporal pattern, assuming a more variety of

covariance structures of the response variable than the traditional growth curve model. Model selection is a crucial issue for the proposed method since it allows fitting a broad range of models and comparing them to each other. A recently developed model fit index, QIC, appears useful for model selection in the proposed method.

The proposed method may be further extended so as to strengthen its data-analytic capability. For instance, reduced-rank restrictions such as $\text{rank}(\mathbf{AB}) \leq \min(D,P)$ (Albert & Kshirsagar, 1993; Reinsel & Velu, 1998, pp. 171-176) may be imposed on the method. The resultant restricted model may offer more parsimonious solutions than the unrestricted one if the rank restrictions are consistent with the data. In the growth curve model, polynomials are typically employed as basis functions. Yet, they appear less efficient for specifying the shape of complex time-varying data (Ramsay, in press). Thus, more diverse kinds of basis functions may also be considered: For instance, the B-spline basis functions may be a good candidate for non-periodic data. Fourier series appears suitable for periodic data (Ramsay & Silverman, 1997). Finally, extended generalized estimating equations (EGEE) (Hall & Severini, 1998) were proposed that combined the original GEE idea with extended quasi-likelihood (McCullagh & Nelder, 1989, p. 349). It appears feasible to adopt EGEE for parameter estimation of the growth curve model. In this case, however, model selection may be more challenging because no formal model fit index such as QIC is yet available. All of these extensions warrant further attention and provide the fodder for future studies.

References

Albert, J. M., & Kshirsagar, A. M. (1993). The reduced-rank growth curve model for discriminant analysis of longitudinal data. *Australian Journal of Statistics*, **35**, 345-357.

Bagozzi, R. P., & Bergami, M. (2002). Antecedents and consequences of organizational identification and the nomological validity of the Bergami-Bagozzi scale. Paper submitted for publication.

Diggle, P. J. (1988). An approach to the analysis of repeated measurements. *Biometrics*, **44**, 959-971.

Duncan, T. E., Duncan, S. C., Alpert, A., Hops, H., Stoolmiller, M., & Muthén, B. (1997). Latent variable modeling of longitudinal and multilevel substance use data. *Multivariate Behavioral Research*, **32**, 275-318.

Duncan, T. E., Duncan, S. C., Hops, H., & Stoolmiller, M.(1995). An analysis of the relationship between parent and adolescent marijuana use via generalized estimating equation methodology. *Multivariate Behavioral Research*, **30**, 317-339.

Grizzle, J. E., & Allen, D. M. (1969). Analysis of growth and dose response curves. *Biometrics*, **25**, 357-381.

Guttman, L. A. (1954). A new approach to factor analysis: The radix. In P. F. Lazarsfeld (Ed.), *Mathematical Thinking in the Social Sciences* (pp. 3-22). New York: Columbia University Press.

Hall, D. B., & Severini, T. A. (1998). Extended generalized estimating equations for clustered data. *Journal of the American Statistical Association*, **93**, 1365-1375.

Jöreskog, K. G. (1970). Estimation and testing of simplex models. *British Journal of Mathematical and Statistical Psychology*, **23**, 121-145.

Khatri, C. G. (1966). A note on a MANOVA model applied to problems in growth curves. *Annals of the Institute of Statistical Mathematics*, **18**, 75-86.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963-974.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22.

MacCallum, R. C., Kim, C., Malarkey, W. B., & Kiecolt-Glaser, J. K. (1997). Studying multivariate change using multilevel models and latent curve models. *Multivariate Behavioral Research*, **32**, 215-253.

McCullagh, P. (1983). Quasi-likelihood functions. *The Annals of Statistics*, **11**, 59-67.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman & Hall.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, **57**, 120-125.

Poffhoff, R. F., & Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, **51**, 313-326.

Ramsay, J. O. (in press). Multilevel modeling of longitudinal and functional data. To appear in D. Moskowitz & S. Hershberger (Eds.), *Modeling Intraindividual Variability with Repeated Measures data: Methods and Applications*. New York: Erlbaum Associates.

Ramsay, J. O., & Silverman, B. W. (1997). *Functional Data Analysis*. New York: Springer-Verlag.

Rao, C. R. (1965). The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika*, **52**, 447-458.

Reinsel, G. C., & Velu, R. P. (1998). *Multivariate Reduced-Rank Regression: Theory and Applications.* New York: Springer-Verlag.

Vonesh, E. F., & Carter, R. L. (1987). Efficient inference for random-coefficient growth curve models with unbalanced data. *Biometrics*, **43**, 617-628.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, **61**, 439-447.

Zeger, S. L, & Liang, K. Y. (1986). The analysis of discrete and continuous longitudinal data. *Biometrics*, **42**, 121-130.

Table 1. Descriptive statistics of alcohol use data measured at four time points.

	Mean	S.D.	Minimum	Maximum
Time 1	2.23	1.04	1	5
Time 2	2.46	1.00	1	5
Time 3	2.65	1.00	1	5
Time 4	2.94	0.94	1	5

Table 2. Summary of fit for various growth curve models with four different covariance structures for the alcohol use data.

Temporal Pattern	Covariance Structure	QIC
Model 1 (Quadratic)	Independence	2293.9
	Exchangeable	2284.4
	AR-1	2284.4
	Unconstrained	2285.3
Model 2 (Linear)	Independence	2290.9
	Exchangeable	2279.5
	AR-1	2274.9
	Unconstrained	2275.8
Model 3 (Stable)	Independence	2454.8
	Exchangeable	2442.4
	AR-1	2438.0
	Unconstrained	2446.8

Table 3. The GEE estimates of **B** in the final model for the alcohol use data with their robust standard errors in the parentheses.

	Initial	Linear
Intercept	.70 (.23)	.20 (.06)
Marital status	-.27 (.10)	.02 (.03)
Family status	.23 (.07)	.01 (.02)
SES	-.07 (.06)	-.02 (.01)
Age	.05 (.06)	.02 (.01)
Gender	.12 (.02)	-.01 (.00)