数理解析研究所講究録
第1603巻 2008年 102-110

Generalized Linear Models Using Trajectories Estimated from a Linear Mixed Model

Nami Maruyama

Division of Biostatistics
Kitasato University Graduate School of Pharmaceutical Sciences

1 Introduction

Longitudinal studies are designed to measure intra-individual change over time. In recent years, longitudinal data has received much attention in many fields such as biomedical research and economics. One reason for the interest stems from the new opportunities provided by longitudinal data to develop predictive models of subsequent outcomes given the current data for an individual. In other words, trajectories, or longitudinally changing patterns of repeated measurements of variables up to a given time t, may afford predictive ability for subsequent observations that are measured after time t, the termination of the trajectory.

A growth curve model based on a linear mixed model helps investigate an overall pattern of change in repeated measurements over time, in other words, trajectories, which can be used to predict subsequent observations. Several works about prediction of the separate outcome variable given current data for an individual by using trajectories by a growth curve model have been reported. Dang et al. (2007) developed a method to use the estimated trajectories of each subject by a bivariate growth curve from longitudinal measures in a Cox proportional hazards model to predict a separate outcome. Other works explored the effects of a measurement error in a time-varying covariate for a mixed model applied to a longitudinal study: Tosteson et al. (1998) used a likelihood-based method of estimation to fit mixed models with measurement errors in covariates. Regression calibration is simple and potentially applicable to any regression model and it tends to be most useful for estimating parameters in GLMs with covariate measurement errors (Carroll et al., 2006).

One challenge to predict subsequent outcome is that the features of the longitudinal profiles are observed only through the longitudinal measurements, which are subject

to measurement error and other variation. Naive implementation by imputing the longitudinal profiles with measurement error yields biased inference, and several methods for reducing this bias have been proposed. For example, Wang et al. (2000) proposed to correct bias with longitudinal covariates with errors in GLMs for binary outcomes and Cox proportional hazards model for censored outcome variables. Although many works have been reported on correcting bias in GLMs, a method of prediction based on GLMs, specifically a logistic regression model, to correct bias from a growth curve model adjusted for other covariates has not been well investigated. In this paper, we investigate a method of prediction by generalized linear models, especially a logistic regression model using trajectories, which are conditional expectations of predictors: In the first step, we get parameter estimates of trajectories which are random effects obtained from a linear mixed model while adjusting other covariates. To predict the subsequent outcome, in the second step, a conditional likelihood approach is applied to correct the parameter estimation errors of trajectories from the first step.

2 Statistical Models

Several modeling approaches have been described in the literature to deal with longitudinal profiles of the data. To this aim, linear mixed models and generalized linear models can be used. In this section, a brief review of these models is described.

2.1 Linear Mixed Model

In mixed effects models, random effects are used to describe the correlation structure in the data and the responses are usually assumed to be independent conditional on the random effects (Verbeke and Molenberghs, 2000; Molenberghs and Verbeke, 2005). The key feature of mixed models is the presence of parameters that vary randomly with the subunits (e.g., persons in longitudinal data; studies in meta-analysis, etc.). For example, linear mixed models are models that incorporate both fixed effects, which are parameters associated with an entire population or with certain repeatable levels of the experimental factors and random effects, which are associated with individual experimental units drawn at random from a population. Linear mixed models are primarily used to describe relationships between a response variable and some covariates in the data that are grouped according to one or more classification factors. By associating common random effects to observations sharing the same level of a classification factor, linear mixed models flexibly represent the covariance structure induced by the data. In some situations linear mixed models are the most plausible models for a particular data structure. Potential advantages are:

- an appropriate covariance pattern model (which directly models a pattern of correlations between observations) leads to more efficient and precise fixed effect estimates and standard errors.
- when the data are hierarchical, their structure can be more directly reflected, leading to a more natural inference.
- the model can provide estimates of the random effects which can be used to distinguish or classify the sub-units.
- linear mixed model offers flexibility in fitting different variance-covariance structures.

A potential disadvantage of linear mixed models is that more distributional assumptions need to be made. Moreover, usually approximations have to be used to estimate certain parameters of the model. As a consequence, conclusions depend on more assumptions, increasing the risk of misspecifying the model and hence biased parameter estimates. Nevertheless, linear mixed models offer a powerful and flexible tool for analysis of longitudinal data.

In longitudinal studies, a growth curve model based on a linear mixed model including two random effects (intercept and slope) which are normally distributed with an independent Gaussian error is probably the most routinely used to study change over time of a quantitative outcome. Therefore, the growth curve model is applied to predict a subsequent outcome in this paper.

2.2 Generalized Linear Models

As a paradigm for a large class of problems in applied statistics, generalized linear models (GLMs) have proved very effective since their introduction by Nelder and Wedderburn (1972). GLMs are a unified class of regression models for discrete and continuous response variables, and have been used routinely in dealing with observational studies.

GLMs have several areas of application ranging from medicine to economics, quality control and sample surveys. Applications of the logistic regression model, expanded with the popularity of case-control designs in epidemiology, now provide a basic tool for epidemiologic investigation of chronic diseases. Probit and logistic models play a key role in all forms of assay experiments. The log-linear model is the cornerstone of modern approaches to the analysis of contingency table data, and has been found particularly useful for medical and social sciences. Poisson regression models are widely employed to study rates of events such as disease outcomes. The complementary log-log model arises in the study of infectious diseases, and more generally, in the analysis of survival data associated with clinical and longitudinal follow-up studies.

Traditionally, the exponential family model adopted for the study of GLMs deals with a linear function of the response variable involving the unknown parameters of interest. This covers most of the experimental situations arising in practice. However, some special members, such as the curved exponential family of distributions, are not covered. Thus, GLMs should be further generalized to include such members.

All known response surface techniques were developed within the framework of linear models under the strong assumptions of normality and equal variances concerning the error distribution. One important area that needs further investigation under the less rigid structure of GLMs is the choice of design. In this paper, we focus on a logistic regression to predict a subsequent outcome in the proposed approaches.

3 Generalized Linear Model with Covariate Measurement Error

In this section, we propose a method of prediction for an outcome variable based on a generalized linear model, specifically a logistic regression model, whose covariates are variables that characterize an individual trajectory. As an individual trajectory contains estimation error, this, in fact, constitutes a measurement error model. The model is fitted in two steps. First, a linear mixed model is fitted to the longitudinal data to estimate the random effect that characterizes the trajectory for each individual while adjusting for other covariates. In the second step, a conditional likelihood approach is applied to account for the estimation error in the trajectory. Prediction of an outcome variable is based on the logistic regression model in the second step.

3.1 The First Step: Trajectories of Longitudinal data

Let W_{ij} denote observations of subject i $(i = 1, \ldots, n)$ at the time point t_j $(j = 1, \ldots, m_i)$. The observation vector of trajectories, $\mathbf{W}_i = (W_{i1}, \ldots, W_{im_i})^t$, is assumed to follow a linear mixed model (Laird and Ware, 1982).

$$\mathbf{W}_i = \mathbf{T}_i \mathbf{U}_i + \mathbf{1} \cdot (\boldsymbol{\beta}^t \mathbf{X}_i) + \boldsymbol{\varepsilon}_i, \tag{3.1}$$

where $\mathbf{T}_i$ is a $m_i \times q$ matrix whose rows consists of $\mathbf{T}_{ij}^t$s; $\mathbf{T}_{ij}$s defined as a $q \times 1$ observed covariate vector which is function of time, $\mathbf{U}_i$ are the $q \times 1$ unobserved vector of subject specific random effects following a normal distribution $\mathrm{N}(\boldsymbol{\xi}, \boldsymbol{\Sigma})$, $\mathbf{1}$ is a $m_i \times 1$ vector whose elements are 1, $\mathbf{X}_i^t$ is $p_1 \times 1$ observed covariate vector with mean $\boldsymbol{\eta}_x$ and variance-covariance matrix $\boldsymbol{\Sigma}_x$, $\boldsymbol{\beta}$ is $p_1 \times 1$ vector containing fixed effects other than times, and $\boldsymbol{\varepsilon}_i$ is a $m_i \times 1$ vector of errors, $(\varepsilon_{i1}, \ldots, \varepsilon_{im_i})^t$ whose elements consist of ε_{ij} following $\mathrm{N}(0, \sigma_\varepsilon^2)$.

We assume that $\mathbf{U}_i$ and $\mathbf{X}_i$ are independent. A typical example is that $\mathbf{T}_{ij}^t = (1, t_{ij})$ with $q = 2$ and components U_{i1} and U_{i2} of $\mathbf{U}_i = (U_{i1}, U_{i2})^t$ can be interpreted as the baseline value and the rate of change of subject i, respectively.

From equation (3.1), the conditional distribution of $\mathbf{U}_i$ given $\mathbf{W}_i$ and $\mathbf{X}_i$ is multivariate normal with mean and variance given by:

$$\begin{aligned}
&\mathrm{E}(\mathbf{U}_i|\mathbf{W}_i, \mathbf{X}_i) = \hat{\mathbf{U}}_i \\
&= \boldsymbol{\xi} + \begin{pmatrix} \boldsymbol{\Sigma}\mathbf{T}_i^t & \mathbf{0}^t \end{pmatrix} \begin{pmatrix} \mathbf{T}_i\boldsymbol{\Sigma}\mathbf{T}_i^t + (\mathbf{1}\cdot\mathbf{1}^t)\cdot\boldsymbol{\beta}^t\boldsymbol{\Sigma}_x\boldsymbol{\beta} + \sigma_\varepsilon^2\mathbf{I} & \mathbf{1}\cdot\boldsymbol{\beta}^t\boldsymbol{\Sigma}_x \\ (\mathbf{1}\cdot\boldsymbol{\beta}^t\boldsymbol{\Sigma}_x)^t & \boldsymbol{\Sigma}_x \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{W}_i - (\mathbf{T}_i\boldsymbol{\xi} + \mathbf{1}\cdot\boldsymbol{\beta}^t\boldsymbol{\eta}_x) \\ \mathbf{X}_i - \boldsymbol{\eta}_x \end{pmatrix} \\
&\mathrm{V}(\mathbf{U}_i|\mathbf{W}_i, \mathbf{X}_i) = \mathbf{A}_i \\
&= \boldsymbol{\Sigma} - \begin{pmatrix} \boldsymbol{\Sigma}\mathbf{T}_i^t & \mathbf{0}^t \end{pmatrix} \begin{pmatrix} \mathbf{T}_i\boldsymbol{\Sigma}\mathbf{T}_i^t + (\mathbf{1}\cdot\mathbf{1}^t)\cdot\boldsymbol{\beta}^t\boldsymbol{\Sigma}_x\boldsymbol{\beta} + \sigma_\varepsilon^2\mathbf{I} & \mathbf{1}\cdot\boldsymbol{\beta}^t\boldsymbol{\Sigma}_x \\ (\mathbf{1}\cdot\boldsymbol{\beta}^t\boldsymbol{\Sigma}_x)^t & \boldsymbol{\Sigma}_x \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{T}_i\boldsymbol{\Sigma} \\ \mathbf{0} \end{pmatrix}.
\end{aligned}$$

3.2 The Second Step: Trajectory Modeling to Predict Subsequent Outcome

Suppose that the subsequent outcomes of interest are responses Y_i which have a distribution in the canonical exponential family (McCullagh and Nelder, 1998), whose mean is given as μ_i, where for a monotonic differential link function $g(\cdot)$. The effects of $(\mathbf{W}_i, \mathbf{X}_i)$ on Y_i are modeled through $(\mathbf{U}_i, \mathbf{Z}_i)$ using the generalized linear model

$$g(\mu_i) = \phi_0 + \mathbf{Z}_i^t\boldsymbol{\phi}_1 + \mathbf{U}_i^t\boldsymbol{\phi}_2,$$

where $\mathbf{Z}_i$ is a $p_2 \times 1$ covariate vector and $\mathbf{U}_i$ is the subject specific random effects from equation (3.1). Wang et al. (2000) proposed a generalized linear model which included longitudinal covariates with a measurement error. We propose a method to use the summary characteristics of trajectories from a growth curve model as predictors in a generalized linear model by extending the method by Wang et al. (2000).

The likelihood function of $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\gamma})$, where $\boldsymbol{\phi} = (\phi_0, \boldsymbol{\phi}_1, \boldsymbol{\phi}_2)$ and $\boldsymbol{\gamma} = (\boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_x, \boldsymbol{\eta}_x, \sigma_\varepsilon^2)$ contributed by the i th subjects is

$$L(Y_i, \mathbf{W}_i; \boldsymbol{\theta}) = L(Y_i|\mathbf{W}_i; \boldsymbol{\theta})L(\mathbf{W}_i; \boldsymbol{\gamma}) \tag{3.2}$$

Therefore, the parameters specifying the assumed models, $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$ are estimated by

maximizing the following log likelihood function.

$$\begin{aligned} \log L(Y, \mathbf{W}; \boldsymbol{\theta}) &= \log L(Y|\mathbf{W}; \boldsymbol{\theta}) + \log L(\mathbf{W}; \boldsymbol{\gamma}) \\ &= \sum_{i=1}^{n} \{\log L(Y_i|\mathbf{W}_i; \boldsymbol{\theta}) + \log L(\mathbf{W}_i; \boldsymbol{\gamma})\}. \end{aligned} \tag{3.3}$$

We consider to maximize the conditional log likelihood $\sum_{i=1}^{n} \log L(Y_i|\mathbf{W}_i; \boldsymbol{\theta})$ with respect to $\boldsymbol{\phi}$ given $\boldsymbol{\gamma}$ and estimate $\boldsymbol{\gamma}$ by maximizing $\sum_{i=1}^{n} \log L(\mathbf{W}_i; \boldsymbol{\gamma})$ through fitting the linear mixed model (3.1).

The true marginal mean for the hierarchical model with normally distributed random effects could be expressed with adjusted values for the regression variables or regression coefficients (Zeger et al., 1988; Breslow and Clayton, 1993). Therefore, $Y_i|\mathbf{W}_i$ follows a generalized linear mixed model:

$$g(\mu_i) = \phi_0 + \mathbf{Z}_i^t \boldsymbol{\phi}_1 + \hat{\mathbf{U}}_i^t \boldsymbol{\phi}_2 + b_i,$$

where $b_i \sim \mathrm{N}(\boldsymbol{\phi}_2^t \mathbf{A}_i \boldsymbol{\phi}_2)$.

Suppose that the subsequent outcome of interest is a binary response Y which follows the logistic regression model $p_i = P(Y = 1|\mathbf{Z}, \mathbf{U}) = H(\phi_0 + \mathbf{Z}_i^t \boldsymbol{\phi}_1 + \mathbf{U}_i^t \boldsymbol{\phi}_2)$, where $H(v) = \{1 + \exp(-v)\}^{-1}$ is the logistic function. It is shown that using the normal approximation to the standard logistic distribution (Johnson et al., 2004; Lin and Breslow, 1999; Wang et al., 2000; Carroll et al., 2006) when some of the predictors, $\mathbf{U}_i$ in this case, are measured with error,

$$\mathrm{logit}(p_i) = \frac{\phi_0 + \mathbf{Z}_i^t \boldsymbol{\phi}_1 + \hat{\mathbf{U}}_i^t \boldsymbol{\phi}_2}{\{1 + (\boldsymbol{\phi}_2^t \mathbf{A}_i \boldsymbol{\phi}_2)/c^2\}^{1/2}}, \tag{3.4}$$

where $c = 15\pi/16\sqrt{3}$. In general, as the denominator in equation (3.4) is greater than 1, estimates of the parameters are slightly attenuated.

By maximizing the log likelihood function (3.3) using the conditional likelihood having covariates with errors, parameters in the second step are estimated. The variance of parameters in equation (3.4) can be estimated by inverting the observed information matrix. Prediction of the outcome variable is based on the logistic regression model in the second step.

4 Discussion

Although generalized linear models with one-time covariates are well understood, little has been done on generalized linear models with longitudinal covariates. We develop

in this paper a framework for generalized linear models with longitudinal covariates. Specifically, we assume that the longitudinal covariates follow a growth curve model based on a linear mixed model and the primary outcome variable depends on the longitudinal covariates through latent subject-specific random effects.

The two-step approach fits the generalized linear model by replacing the unobserved random effects with their estimators obtained by fitting individual linear regression models using each subject's data from the first step. The conditional likelihood approach estimates the regression coefficients by maximizing the conditional distribution of the outcome variable given the observed longitudinal covariates.

To relate trajectories and subsequent outcomes, a two-step approach can be taken as well as a single-step joint distribution model. The two-step approach is a simple way to assess causal relationship between outcomes of different stages. However, one disadvantage of a two-step model is its bias when covariates are subject specific estimators obtained from the first step. Several works investigated the bias of estimates in GLMs obtained in the second step and proposed methods to reduce the bias (Dang et al., 2007; Wang et al., 2000). When trajectories from a growth curve model are assessed, covariates affecting trajectories should be taken into account; however, most of the existing works do not take into account the effects of covariates on the trajectories. Trajectories may be affected by other covariates, such as age, gender, or some prognostic factors; hence we propose a conditional likelihood approach to account for these errors and correct the bias to incorporate the effects of covariates.

Although a full likelihood approach could also be use to reduce bias due to estimation errors, the conditional likelihood approach is numerically more practical because it utilizes estimates of trajectories from a linear mixed model in the first step.

There are several directions for further research on this topic. First, further research should be conducted to assess the validity of the proposed models through simulation. Also, implications of misspecification of the models should be studied. An important aspect in this regard would be the development of procedures for model diagnostics.

As Dang et al. (2007) pointed out, when we use the estimated random effects as predictors for another model, the normality assumption is very important but difficult to test. Further research on formal methods for assessing possible departures from normal random effects would be useful. Li et al. (2004) proposed estimators for the generalized linear model parameters that require no assumptions on the random effects. We could investigate semiparametric estimators as proposed by Li et al. (2004) also incorporating covariates which affects trajectories. The two-step approach could be easily extended to handle various parametric models in the first step such as non-linear mixed models or multivariate longitudinal measurement may be entertained; therefore, bias caused by the two-step approach which uses non-linear mixed models should also be investigated.

References

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of American Statistical Association*, 88(421):9–25.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models*. Chapman & Hall, New York, second edition.

Dang, Q., Mazumdar, S., Anderson, S. J., Houck, P. R., and Reynolds, C. F. (2007). Using trajectories from a bivariate growth curve as predictors in a cox regression model. *Statistics in Medicine*, 26(4):800–811.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (2004). *Continuous univariate distributions*, volume 1. John Wiley & Sons, New York, second edition.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38:936–937.

Li, E., Zhang, D., and Davidian, M. (2004). Conditional estimation for generalized linear models when covariates are subject-specific parameters in a mixed model for longitudinal measurements. *Biometrics*, 60:1–7.

Lin, X. and Breslow, N. E. (1999). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of American Statistical Association*, 91:1007–1016.

McCullagh, P. and Nelder, J. A. (1998). *Generalized linear models*. Chapman & Hall, New York, second edition.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer, New York.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of Royal Statistical Society, Series A*, 135:370–384.

Tosteson, T. D., Buonaccorsi, J. P., and Demidenko, E. (1998). Covariate measurement error and the estimation of random effect parameters in a mixed model for longitudinal data. *Statistics in Medicine*, 17:1959–1971.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer, New York.

Wang, C. Y., Wang, N., and Wang, S. (2000). Regression analysis when covariates are regression parameters of a random effects model for observed longitudinal measurements. *Biometrics*, 56:487–495.

Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 44:1049–1060.