Communications of the Korean Statistical Society Vol. 15, No. 6, 2008, pp. 969–976

A General Mixed Linear Model with Left-Censored Data

Il Do Ha[1)]

Abstract

Mixed linear models have been widely used in various correlated data including multivariate survival data. In this paper we extend hierarchical-likelihood(h-likelihood) approach for mixed linear models with right censored data to that for left censored data. We also allow a general random-effect structure and propose the estimation procedure. The proposed method is illustrated using a numerical data set and is also compared with marginal likelihood method.

Keywords: H-likelihood; left censoring; marginal likelihood; mixed linear models; random effects.

1. Introduction

In biomedical research, correlated survival data in the form of repeated, recurrent or multiple event times by clusters including subjects are frequently encountered (Hougaard, 2000; Ha *et al.*, 2001). The correlation can be analyzed via random effects (Laird and Ware, 1982; Hougaard, 2000; Ha *et al.*, 2002). However, the data are usually censored due to limited period of observation, which can lead to a complicated inference. In almost cases, right censoring is common, but left censoring sometimes occurs. In this paper we are interested in case of left censoring, with correlation between the event times. An observation is said to be *left censored* if the exact value of the observation is unknown, but is known only that the observation is smaller than some given value. For example, in a study of age at the onset of puberty in female rats from the same litter, the records may start after puberty has already occurred in some individuals. In such a case, the event of interest is not observed and the record is left censored and can be also correlated because of data from the same litter (Carriquiry *et al.*, 1987). As another example, there occurs a left censoring due to the lower limit of detection in a longitudinal study with repeated measures of human immunodeficiency virus(HIV) viral load, the primary measure of HIV infection (Hughes, 1999; Jacqmin-Gadda *et al.*, 2000).

For the analysis of left or right censored correlated data, mixed linear models(MLMs) have been used as an alternative of frailty hazard models (Hougaard, 2000), in which the random effect acts linearly on each individual's survival time, thus making the interpretation of the fixed effects easier than in the frailty models (Carriquiry *et al.*, 1987; Klein *et al.*, 1999; Jacqmin-Gadda *et al.*, 2000). In particular, the fixed-effect estimates in MLMs are relatively robust against the violations of the model assumptions such as normality

1) Professor, Department of Asset Management, Daegu Haany University, Gyeongsan 712-715, Korea. E-mail: idha@dhu.ac.kr

of random-effect distribution (Ha *et al.*, 2002). For the inferences of MLMs with the left-censored data, several authors have proposed marginal likelihood(ML) methods such as Monte Carlo EM (Hughes, 1999) because of the required intractable integrations. Recently, Thiébaut and Jacqmin-Gadda (2004) also presented the use of SAS PROC NLMIXED based upon Gauss-Hermite quadrature(GHQ). However, the ML methods are still computationally intensive and not practical, particularly for mult-component MLMs including mult-level or crossed structures (Gueorguieva, 2001; Ha and Lee, 2005).

Lee and Nelder (1996, 2001) have introduced h-likelihood which avoids the intractable integrals necessary to obtain the marginal likelihood. They have showed that the h-likelihood provides a statistically efficient procedure for models(*e.g.* hierarchical generalized linear models) with various random-effect structures. Ha *et al.* (2002) and Ha and Lee (2005) have also developed h-likelihood procedures for fitting one random component and multi-component MLMs with correlated survival data, respectively. Furthermore, Ha *et al.* (2007b) proposed the h-likelihood method on genetic MLMs for twin survival data under left truncation. In particular, for the MLMs the h-likelihood provides a conceptually simple, numerically efficient and reliable inferential procedure. However, the h-likelihood inferences have been studied in the presence of right censoring only. In this paper we extend the h-likelihood procedure of MLM under right censoring to that of left censoring. Here we allow a correlation between random effects. The proposed method is also compared with marginal likelihood method (Thiébaut and Jacqmin-Gadda, 2004) using SAS NLMIXED.

In Section 2 we describe a general MLM under left censoring. In Section 3 we propose the h-likelihood estimation procedure, leading to the use of the pseudo-variable. The proposed method is demonstrated using a numerical data set in Section 4, followed by discussion in Section 5.

2. The Model

Let T_{ij} $(i = 1, \ldots, q,\ j = 1, \ldots, n_i,\ n = \sum_i n_i)$ be the survival time for the j^{th} observation of the i^{th} cluster and C_{ij} be the corresponding left-censoring time. The observable random variables are

$$Y_{ij} = \max(T_{ij},\, C_{ij}) \quad \text{and} \quad \delta_{ij} = I(T_{ij} \geq C_{ij}),$$

where $I(\cdot)$ is the indicator function.

For T_{ij} we consider a general MLM, which allows the correlation between random effects: for $i = 1, \ldots, q$ and $j = 1, \ldots, n_i$,

$$T_{ij} = x_{ij}^t \beta + z_{ij}^t U_i + \epsilon_{ij}, \tag{2.1}$$

where $x_{ij} = (1,\, x_{ij1}, \ldots, x_{ijp})^t$ is a vector of fixed covariates and β is a $(p+1) \times 1$ vector of fixed effects. Here $U_i = (U_{i1}, \ldots, U_{id})^t$ is a d-dimensional vector of random effects from the i^{th} cluster, which is associated with a vector of random covariates z_{ij}. Note that $U_i \sim N(0,\, A_i)$ and $\epsilon_{ij} \sim N(0,\, \sigma_\epsilon^2)$ are assumed to be independent and that the covariance matrix $A_i = A_i(\theta)$ depends on θ denoting a vector of unknown parameters. For example, in case of bivariate normal distribution($d = 2$) $A_i(\theta)$ has three parameters with $\theta_1 = \sigma_1^2$, $\theta_2 = \sigma_2^2$ and $\theta_3 = \sigma_{12}$; as a special case $A_i(\theta) = \mathrm{diag}(\theta_1,\, \theta_2)$ becomes a diagonal matrix if $\theta_3 = 0$(*i.e.* the correlation is zero).

In model (2.1), z_{ij} is often a subset of x_{ij}. If $z_{ij} = 1$ for all i, j and $U_i = U_{i0}$, then the model (2.1) with $z_{ij}^t U_i = U_{i0}$ reduces to a random intercept MLM (Ha *et al.*, 2002). Furthermore, if $z_{ij} = (1, x_{ij1})^t$ and $U_i = (U_{i0}, U_{i1})^t$, then $z_{ij}^t U_i = U_{i0} + U_{i1} x_{ij1}$ and the corresponding model becomes a random intercept (U_{i0}) and slope (U_{i1}) MLM (Thiébaut and Jacqmin-Gadda, 2004).

Notice that the T_{ij} can be expressed on some suitably transformed scale, *e.g.* $\log(T_{ij})$. If the log-transformation is used, the MLM (2.1) becomes an accelerated failure-time model with random effects.

3. Estimation Procedure

Following Ha *et al.* (2001, 2002), the h-likelihood for the model (2.1), denoted by h, is defined by

$$h = h(\beta, \theta, \sigma_\epsilon^2) = \sum_{ij} \ell_{1ij} + \sum_i \ell_{2i}, \tag{3.1}$$

where

$$\ell_{1ij} = \ell_{1ij}(\beta, \sigma_\epsilon^2; y_{ij}, \delta_{ij} \,|\, u_i) = -\delta_{ij} \frac{\log(2\pi\sigma_\epsilon^2) + (m_{ij})^2}{2} + (1 - \delta_{ij}) \log\{\Phi(m_{ij})\}$$

is the logarithm of the conditional density function for Y_{ij} and δ_{ij} given $U_i = u_i$ and

$$\ell_{2i} = \ell_{2i}(\theta; u_i) = -\frac{1}{2}\{\log\det(2\pi A_i(\theta))\} - \frac{1}{2} u_i^t A_i(\theta)^{-1} u_i$$

is the logarithm of the density function for U_i.

Note here that

$$E(Y_{ij} \,|\, U_i = u_i) \neq \mu_{ij},$$

where $\mu_{ij} = E(T_{ij} \,|\, U_i = u_i) = x_{ij}^t \beta + z_{ij}^t u_i$. Following Ha *et al.* (2002) and Ha and Lee (2005), we can show that the h-likelihood method is equivalent to the use of pseudo-responses y_{ij}^*, given by

$$\begin{aligned} y_{ij}^* &= E(T_{ij} \,|\, Y_{ij} = y_{ij}, \delta_{ij}, U_i = u_i) \\ &= y_{ij}\delta_{ij} + B_{ij}(1 - \delta_{ij}), \end{aligned} \tag{3.2}$$

which is an extension of the pseudo-responses under right censoring to left censoring. Here, $B_{ij} = E(T_{ij} \,|\, T_{ij} > y_{ij}, U_i = u_i) = \mu_{ij} - \sigma_\epsilon V(m_{ij})$, $V(\cdot) = \phi(\cdot)/\Phi(\cdot)$, ϕ and Φ are the density and cumulative distribution functions for $N(0, 1)$, respectively and $m_{ij} = (y_{ij} - \mu_{ij})/\sigma_\epsilon$. It can be shown that

$$E(y_{ij}^* \,|\, U_i = u_i) = \mu_{ij}.$$

Thus, the h-likelihood method implicitly applies the EM-type algorithm to the h-likelihood procedure (Ha and Lee, 2005).

3.1. Estimation of fixed and random effects

Given the dispersion components $\psi = (\sigma_\epsilon^2, \theta^t)^t$, the maximum h-likelihood estimators(MHLEs) of $\tau = (\beta^t, u^t)^t$ with $u = (u_1^t, \ldots, u_q^t)^t$ are obtained by solving

$$\frac{\partial h}{\partial \beta_k} = \frac{1}{\sigma_\epsilon} \sum_{ij} \{\delta_{ij} m_{ij} - (1 - \delta_{ij}) V(m_{ij})\} x_{ijk} = 0, \quad (k = 1, \ldots, p+1) \tag{3.3}$$

and

$$\frac{\partial h}{\partial u_i} = \frac{1}{\sigma_\epsilon} \sum_{j} \{\delta_{ij} m_{ij} - (1 - \delta_{ij}) V(m_{ij})\} - A_i^{-1} u_i = 0, \quad (i = 1, \ldots, q). \tag{3.4}$$

Substituting (3.2) into the two MHL equations (3.3) and (3.4) reduces them, respectively, to

$$\frac{1}{\sigma_\epsilon^2} \sum_{ij} (y_{ij}^* - \mu_{ij}) x_{ijk} = 0, \quad (k = 1, \ldots, p+1)$$

and

$$\frac{1}{\sigma_\epsilon^2} \sum_{j} (y_{ij}^* - \mu_{ij}) \quad - A_i^{-1} u_i = 0, \quad (i = 1, \ldots, q).$$

Thus, the MHLEs $\widehat{\tau} = (\widehat{\beta}^t, \widehat{u}^t)^t$ given θ and y^* are obtained by solving Henderson's (1975) mixed-model equations iteratively with pseudo-response variables y^*:

$$\begin{pmatrix} X^t X & X^t Z \\ Z^t X & Z^t Z + \Lambda \end{pmatrix} \begin{pmatrix} \widehat{\beta} \\ \widehat{u} \end{pmatrix} = \begin{pmatrix} X^t y^* \\ Z^t y^* \end{pmatrix}, \tag{3.5}$$

where X and Z is the $n \times (p+1)$ and $n \times q$ model matrices of x_{ij} and z_{ij}, respectively and y^* is the $n \times 1$ vector with ij^{th} element y_{ij}^* and $\Lambda = \sigma_\epsilon^2 A^{-1}$ with the $q \times q$ block diagonal $A = \text{blockdiag}(A_1, \ldots, A_q)$.

From (3.3) and (3.4) the asymptotic covariance matrix (Lee and Nelder, 1996; Ha *et al.*, 2002) for $\widehat{\tau} - \tau$ is given by H^{*-1} with

$$H^* = -\frac{\partial^2 h}{\partial \tau^2} = \frac{1}{\sigma_\epsilon^2} H, \tag{3.6}$$

where

$$H = \begin{pmatrix} X^t W X & X^t W Z \\ Z^t W X & Z^t W Z + \Lambda \end{pmatrix}.$$

Here, $W = \text{diag}(w_{ij})$ is the $n \times n$ diagonal matrix with the ij^{th} weight element $w_{ij} = \delta_{ij} - (1 - \delta_{ij}) \xi(m_{ij})$ and $\xi(m_{ij}) = -V(m_{ij})\{V(m_{ij}) + m_{ij}\}$. So, the upper left-hand corner of H^{*-1} in (3.6) gives the variance matrix of $\widehat{\beta}$:

$$\text{var}(\widehat{\beta}) = \sigma_\epsilon^2 (X^t \Sigma^{-1} X)^{-1},$$

where $\Sigma = W^{-1} + Z \Lambda^{-1} Z^t$.

3.2. Estimation of dispersion parameters

For the estimation of the dispersion parameters $\psi = (\sigma_\epsilon^2, \theta^t)^t$, we use Lee and Nelder's (2001) adjusted profile h-likelihood, defined by

$$p_\tau(h) = \left[h - \frac{1}{2} \log \det \left\{ \frac{D(h, \tau)}{(2\pi)} \right\} \right] \Bigg|_{\tau=\widehat{\tau}}, \tag{3.7}$$

where h and $D(h, \tau) = -\partial^2 h/\partial \tau^2 = H^*$ are given in (3.1) and (3.6), respectively and $\widehat{\tau} = \widehat{\tau}(\psi) = (\widehat{\beta}^t(\psi), \widehat{u}^t(\psi))^t$. Note that $p_\tau(h)$ is an extension of restricted likelihood (Patterson and Thompson, 1971) of MLM without censoring to that with censoring (Ha *et al.*, 2002). The restricted maximum likelihood(REML) estimator for ψ are obtained by solving iteratively

$$\frac{\partial p_\tau(h)}{\partial \psi} = 0.$$

Firstly, $\partial p_\tau(h)/\partial \sigma_\epsilon^2 = 0$ gives the REML estimator for σ_ϵ^2:

$$\widehat{\sigma_\epsilon^2} = \frac{(y^* - \widehat{\mu})^t (y^* - \widehat{\mu})}{n_0 - (p + q - \gamma_0)},$$

where $\widehat{\mu} = X\widehat{\beta} + Z\widehat{u}$, $n_0 = \sum_{ij} w_{ij}$ and $\gamma_0 = \sigma_\epsilon^2 \text{trace}\{H^{-1}(\partial H/\partial \sigma_\epsilon^2)\}$. Secondly, the REML estimators for θ are obtained by solving the estimation equations

$$\frac{\partial p_\tau(h)}{\partial \theta} = -\frac{1}{2}\text{trace}\left\{ A^{-1} \left(\frac{\partial A}{\partial \theta} \right) \right\} + \frac{1}{2} \widehat{u}^t \left(\frac{\partial A^{-1}}{\partial \theta} \right) \widehat{u} - \frac{1}{2}\text{trace}\left\{ \mathrm{H}^{-1} \left(\frac{\partial \mathrm{H}}{\partial \theta} \right) \right\} = 0,$$

where $\partial A^{-1}/\partial \theta = -A^{-1}(\partial A/\partial \theta)A^{-1}$ and for the solutions we use the Newton-Raphson method. Appendix II of Ha *et al.* (2002) gives the formulation for the $\partial H/\partial \psi$. Here the $\partial \widehat{u}/\partial \psi$ term should be allowed (Ha and Lee, 2005).

Note that since we cannot observe all the y_{ij}^*'s due to censoring, we substitute estimates, say $\widehat{y_{ij}^*}$, for them in each iteration. For the corresponding fitting algorithm we follow that of Ha *et al.* (2002).

Lee and Nelder (2001) showed that in MLMs without censoring the h-likelihood method provides the maximum likelihood estimators for fixed effects (using Henderson's (1975) equations) and the REML estimators for dispersion parameters. Now we see that for MLMs with left-censoring it implicitly implements an EM-type algorithm by replacing unobserved responses T_{ij} with $E(T_{ij} \,|\, Y_{ij} = y_{ij}, \delta_{ij}, U_i = u_i)$ in the estimating equations (3.5). With the use of h-likelihood the numerically difficult E-step or integration is avoided by automatically imputing the censored responses to y_{ij}^*.

4. Numerical Example

The proposed method is illustrated using a numerical example, based upon a simulated data set from Lyles *et al.* (2000) and Thiébaut and Jacqmin-Gadda (2004). The data($n = 250$ with $q = 50$ and $n_i = 5$) are generated from a random intercept and slope MLM as a special case of the model (2.1):

$$T_{ij} = (\beta_0 + U_{i0}) + (\beta_1 + U_{i1}) x_{ij1} + \epsilon_{ij}, \tag{4.1}$$

Table 4.1: Results on the estimation of parameters under three MLMs

			Factor level						
Model	$\hat{\beta}_0$(SE)	$\hat{\beta}_1$(SE)	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\sigma}_{12}$	$\hat{\sigma}_\epsilon^2$	$p_\tau(h)$	p_T	AIC
True	3.00	0.50	0.50	0.100	−0.10	0.20			
M1	2.95(0.12)	0.50(0.049)	0.54	–	–	0.54	−255.26	2	514.52
M2	2.96(0.12)	0.49(0.057)	0.53	0.058	–	0.24	−252.64	3	511.28
M3	2.95(0.13)	0.50(0.062)	0.67	0.092	−0.11	0.23	−250.83	4	509.66
M3*	2.94(0.13)	0.51(0.062)	0.66	0.089	−0.11	0.23			

Note: SE, the estimated standard error; M3*, results (Thiébaut and Jacqmin-Gadda, 2004) of M3 using SAS PROC NLMIXED

where x_{ij1} is the j^{th} repeated time of the i^{th} subject, $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ and

$$\begin{pmatrix} U_{i0} \\ U_{i1} \end{pmatrix} \sim N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right\}.$$

Here, the true parameters are $\beta_0 = 3$, $\beta_1 = 0.5$, $\theta = (\sigma_1^2, \sigma_2^2, \sigma_{12})^t = (0.5, 0.1, -0.1)^t$ and $\sigma_\epsilon^2 = 0.2$. In the simulated data set, the left-censored observations were replaced by the value(*i.e.* 2.5) of the threshold, leading to 15.2% left-censoring rate. The data are available at http://www.blackwellpublishing.com/rss/Volumes/Cv49p4.htm.
We considered the three submodels of (4.1):

M1 : one-component (U_{i0}) MLM,

M2 : two-components (U_{i0}, U_{i1}) MLM with $\sigma_{12} = 0$,

M3 : two-components (U_{i0}, U_{i1}) MLM with $\sigma_{12} \neq 0$.

For the model fitting we used SAS/IML based on the h-likelihood procedure in Section 3. The results are presented in Table 4.1. The two nested models(M1 and M2) ignoring one random component or covariance(*i.e.* correlation) work well for the estimation of fixed effects, β_0 and β_1. However, in M1 and M2 the standard errors for estimator of fixed slope β_1 are underestimated compared to that of M3 and dispersion-parameter estimates are biased; in particular, the larger biases occur from σ_ϵ^2 in M1 and σ_2^2 in M2, respectively. Overall, the estimates in true model(M3) perform well. For the M3 the h-likelihood estimation results are very similar to the ML results (Thiébaut and Jacqmin-Gadda, 2004) using SAS PROC NLMIXED.

Furthermore, we tried to confirm the selection of true model(M3) among the three models considered. For this we used the Akaike information criterion(AIC) method (Ha *et al.*, 2007a) based on the restricted likelihood $p_\tau(h)$ in (3.7), given by

$$\text{AIC}(T_d) = T_d + 2p_T,$$

where $T_d = -2p_\tau(h)$ and p_T is the number of dispersion parameters. We select the model with smallest AIC as the best model among these models. Note here that if the AIC difference is larger than $1 \sim 2$ it is considered to be significant and that if the difference is less than 1 it is not (Sakamoto *et al.*, 1986; Ha *et al.*, 2007a). The AIC difference between M1 and M2(M2 and M3) is 3.24(1.62), respectively. Thus, under this empirical criterion we find that the AIC selects M3 as the final model.

5. Discussion

We have showed that the h-likelihood procedure for MLM with right censoring can be straightforwardly extended to the left-censoring. We have also found via an empirical study that the proposed method performs well. In particular, we have confirmed that the h-likelihood and marginal likelihood results are about the same. However, the simulation study in Section 4 is somewhat limited because it uses a simple setting such as a sample size and a parameter value. Thus, the performance of proposed method may be potential. As a further work, it is required to compare both results using simulated data from various sample sizes and parameter values.

The h-likelihood method can be easily applied to various random-effect models because it avoids an intractable integration. However, the use of ML method can be limited; for example, for multi-component models with nested or crossed random effects the ML requires a numerically difficult integration, while the h-likelihood does not (Thiébaut and Jacqmin-Gadda, 2004; Ha and Lee, 2005). Furthermore, the h-likelihood method is also useful in selecting a suitable model among a set of candidate models, as shown in Section 4: see also Ha *et al.* (2007a).

Though not reported here, we have found that our method can be also extended to doubly censored data (*i.e.* both left and right censored data, Turnbull, 1974; Hughes, 1999).

References

Carriquiry, A. L. and Gianola, D. and Fernando, R. L. (1987). Mixed-model analysis of a censored normal distribution with reference to animal breeding, *Biometrics,* **43**, 929–939.

Gueorguieva, R. (2001). A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family, *Statistical Modelling,* **1**, 177–193.

Ha, I. D. and Lee, Y. (2005). Multilevel mixed linear models for survival data, *Lifetime data Analysis,* **11**, 131–142.

Ha, I. D., Lee, Y. and MacKenzie, G. (2007a). Model selection for multi-component frailty models, *Statistics in Medicine,* **26**, 4790–4807.

Ha, I. D., Lee, Y. and Pawitan, Y. (2007b). Genetic mixed linear models for twin survival data, *Behavior Genetics,* **37**, 621–630.

Ha, I. D., Lee, Y. and Song, J. K. (2001). Hierarchical likelihood approach for frailty models, *Biometrika,* **88**, 233–243.

Ha, I. D., Lee, Y. and Song, J. K. (2002). Hierarchical-likelihood approach for mixed linear models with censored data, *Lifetime Data Analysis,* **8**, 163–176.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model, *Biometrics,* **31**, 423–447.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data,* Springer-Verlag, New York.

Hughes, J. P. (1999). Mixed effects models with censored data with application to HIV RNA levels, *Biometrics,* **55**, 625–629.

Jacqmin-Gadda, H., Thiébaut, R., Chêne, G. and Commenges, D. (2000). Analysis of left-censored longitudinal data with application to viral load in HIV infection, *Biostatistics,* **1**, 355–368.

Klein, J. P., Pelz, C. and Zhang, M. J. (1999). Modelling random effects for censored data by a multivariate normal regression model, *Biometrics,* **55**, 497–506.

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

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion), *Journal of the Royal Statistical Society, Series B,* **58**, 619–678.

Lee, Y. and Nelder, J. A. (2001). Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions, *Biometrika,* **88**, 987–1006.

Lyles, R. H., Lyles, C. M. and Talyor, D. J. (2000). Random regression models for human immunodeficiency virus ribonucleic acid data subject to left censoring and informative drop-outs, *Journal of the Royal Statistical Society: Series C (Applied Statistics),* **49**, 485–497.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal, *Biometrika,* **58**, 545–554.

Sakamoto, Y., Ishiguro, M. and Kitagawa, G. (1986). *Akaike Information Criterion Statistics,* KTK Scientific Publisher, Tokyo.

Thiébaut, R. and Jacqmin-Gadda, H. (2004). Mixed models for longitudinal left-censored repeated measures, *Computer Methods and Programs in Biomedicine,* **74**, 255–260.

Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data, *Journal of the American Statistical Association,* **69**, 169–173.

[Received August 2008, Accepted October 2008]