The Korean Communications
in Statistics Vol. 12 No. 3, 2005
pp. 759-771

Genetic Mixed Effects Models for Twin Survival Data[1)]

Il Do Ha[2)], Maengseok Noh[3)], and Sangchul Yoon[4)]

Abstract

Twin studies are one of the most widely used methods for quantifying the influence of genetic and environmental factors on some traits such as a life span or a disease. In this paper we propose a genetic mixed linear model for twin survival time data, which allows us to separate the genetic component from the environmental component. Inferences are based upon the hierarchical likelihood (h-likelihood), which provides a statistically efficient and simple unified framework for various random-effect models. We also propose a simple and fast computation method for analyzing a large data set on twin survival study. The new method is illustrated to the survival data in Swedish Twin Registry. A simulation study is carried out to evaluate the performance.

Keywords: Environment effect; Genetic effect; Hierarchical likelihood; Random effects

1. Introduction

Twin studies are useful for judging whether some trait such as a life span or a disease is hereditary. It is interesting to estimate the relative importance of genetic and environmental contributions to the variation of such trait. For this the data on MZ (monozygotic) and DZ (dizygotic) twins are frequently used (Neal and Cardon, 1992) and they have been analyzed using random-effect models, which allow to separate the effects of genetic and environment. In this paper we are interested in genetic analysis using correlated survival time data on the life spans of twins. For the analysis frailty models have been often used: see for example Yashine et al. (1999) and Hougaard (2000). Here, the frailties (or random effects) act multiplicatively on the individual hazard rate.

As an alternative to frailty models, mixed linear models. i.e., mixed effects models (MEMs) have been proposed, in which the random effects act linearly on the individual survival time.

1) This work was supported by Korea Research Foundation Grant (KRF-2003-002-C00045).
2) Associate Professor, Department of Asset Management, Daegu Haany University, Gyeongsan, 712-715, Korea. Email: idha@dhu.ac.kr
3) Postdoctoral, Department of Statistics, Seoul National University, Seoul, 151-742, Korea
4) Lecturer, Department of Computer and Information Science, Daegu Haany University, Gyeongsan, 712-715, Korea.

MEMs have a number of advantages (Lambert et al., 2004; Ha and Lee, 2005a). For example, the estimates from MEMs are robust against various misspecifications about the model assumptions such as neglected covariates or misspecification of the distribution for random effects, while in frailty models (Agresti et al., 2004) they are relatively not. In this paper we propose a genetic MEM for analyzing twin survival data. The proposed model allows to separate the genetic component from the environmental component. However, MEMs have received relatively less attention in the analysis of correlated survival data because of intractable integration required to obtain the marginal likelihood. The h-likelihood avoids such difficulties, giving a statistically efficient and simple unified framework for various random-effect models (Lee and Nelder, 1996; Ha et al., 2002).

The paper is organized as follows. In Section 2 we briefly describe survival data in the Swedish Twin Register. In Section 3 we propose a genetic MEM for the twin survival data. In Section 4 we develop a new h-likelihood procedure, leading to a simple and fast computation for analyzing the MEM with the large twin survival data. In Section 5 the proposed method is applied to the analysis of twin survival data, followed by a numerical study for the performance in Section 6. Finally, some technical details are given in Appendix.

2. The Swedish Twin Survival Data

The Swedish Twin Registry is currently the largest population-based twin registry in the world and includes informations (e.g., life-span, diseases) on twins born in Sweden since 1886.

<Table 1> Survival data in the Swedish Twin Registry, born since 1886.

Number	Pairid	Zygalg	Birthday	Dead	Death.date	Eff.date	Sex
1	11	2	06JAN1900	1	03JAN1987	·	2
2	11	2	06JAN1900	1	15DEC1990	·	2
3	12	2	07JAN1900	1	23DEC1982	·	2
4	12	2	07JAN1900	1	20FEB1994	31AUG1997	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
17	21	1	01JAN1926	0	·	30AUG1997	1
18	21	1	01JAN1926	0	·	30AUG1997	1
19	22	4	01JAN1926	0	·	21JUN2000	1
20	22	4	01JAN1926	1	15MAY1991	·	2
21	23	2	01JAN1926	0	·	03JAN2002	1
22	23	2	01JAN1926	1	13MAR1993	·	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
78207	244783	2	31DEC1958	0	·	06JUN2000	1
78208	244783	2	31DEC1958	0	·	02FEB2001	1
78209	244784	4	31DEC1958	0	·	21MAY1999	1
78210	244784	4	31DEC1958	0	·	30AUG1997	2

Number, number of twins; Pairid, ID of twin pairs; Zygalg (1=MZ, 2=DZ with the same gender, 4=DZ with the opposite gender); Dead (1=dead, 0=alive); Death.date, the date of death; Eff.date, the latest follow-up date; Sex (1=male, 2=female).

This data set was recently provided by Professor Yudi Pawitan in Karolinska Institutet of Sweden: see the website http://www.meb.ki.se. In Table 1 we briefly present the structure of survival data on life-span in the Twin Registry. The survival time is defined as the time to death (i.e., life-span time), measured in years (ages); it is calculated as (the date of death-birthday)/365. If an individual is still alive at the end of follow-up, this life-span datum is also right censored; in this case the date of death is replaced by the latest follow-up date.

<Table 2> Composition of the Swedish twin survival data by sex, zygosity, and censoring status: the old cohort

Data	One censored	Both censored	None censored	Total (pairs)
Males				
MZ	313	159	1174	1646
DZ	620	258	2074	2952
Females				
MZ	450	396	1161	2007
DZ	931	686	2227	3844
Total	2314	1499	6636	10449

The survival data in Table 1 are represented by three different age cohorts, old, middle and young cohorts. In this paper we consider the old cohort: see also Yashin et al. (1999). This cohort consists of all same-sexed pairs born between 1886 and 1925. The data used in the analyses are summarized in Table 2, which are categorized according to the censoring status. Each data set is in the range of a low censoring rate, about 20%~30%. For example, the censoring rate of male MZ twins is 19%, calculated by (313+2×159)/(2×1646). The information in Table 2 shows that there are more female than male twins, which may be explained by the longer female life-span. The ratio of MZ to DZ twin pairs shows about 1:2 (i.e., MZ=3653 : DZ=6796), which confirms the mention by Sham (1998, pp. 189).

3. The Model

In this section we model a direct relationship between twin survival time and covariates including observed or unobserved factors. Let T_{ij} be the survival time (e.g. age at death) for the jth member of the ith twin pair. Let g_{ij} and c_{ij} be the random-genetic effect and the common childhood random-environment effect for the jth individual in the ith twin pair, respectively. For the modelling of skewed data from T_{ij} we use $\log T_{ij}$ as the responses: see also Ha et al. (2002). Thus, we consider the MEM with two random effects: for $i = 1, \cdots, q$ and $j = 1, 2$,

$$\log T_{ij} = \boldsymbol{x}_{ij}^{T}\boldsymbol{\beta} + g_{ij} + c_{ij} + \epsilon_{ij} \ , \tag{1}$$

where $\boldsymbol{x}_{ij}=(x_{ij1},\cdots,x_{ijp})^T$ is a vector of fixed covariates, β is a $p\times 1$ vector of fixed effects, and $g_{ij}\sim N(0,\sigma_g^2)$, $c_{ij}\sim N(0,\sigma_c^2)$ and $\epsilon_{ij}\sim N(0,\sigma_\epsilon^2)$ are mutually independent error components. Note here that between-pair genetic (or environment) effects are independent, but within-pair values are not. Following Sham (1998, pp. 189) and Pawitan et al. (2004), if the ith twin pair is MZ (denoted by MZ_i), it is assumed that

$$\text{corr}\ (g_{i1},g_{i2})=1 \quad \text{and} \quad \text{corr}\ (c_{i1},c_{i2})=1$$

and if it is DZ (denoted by DZ_i),

$$\text{corr}\ (g_{i1},g_{i2})=0.5 \quad \text{and} \quad \text{corr}\ (c_{i1},c_{i2})=1.$$

The discrepancy in genetic correlation between MZ and DZ twins allows us to separate the genetic from the common environmental factor (Pawitan et al., 2004). Let $v_{ij}=g_{ij}+c_{i0}$ for $j=1,2$, where $c_{i0}(=c_{i1}=c_{i2})$ denotes the common environmental effect for the two individuals of the ith twin pair. Then we have that

$$\rho=\text{corr}\ (v_{i1},v_{i2})=\frac{\text{cov}\ (g_{i1},g_{i2})+\sigma_c^2}{\sigma_g^2+\sigma_c^2}.$$

Note that $\rho=1$ for MZ_i and $\rho=(0.5\sigma_g^2+\sigma_c^2)/(\sigma_g^2+\sigma_c^2)\in[0.5,1.0]$ for DZ_i. For the purpose of interpretation it is convenient to define the quantity

$$h_g^2=\frac{\sigma_g^2}{\sigma_g^2+\sigma_c^2+\sigma_\epsilon^2}$$

known as heritability. This concept was introduced in order to measure the importance of genetics in relation to other factors in causing the variability of a trait in a population (Sham, 1998, pp. 212).

In Appendix we show that the two random-effects MEM (1) can be written as a single random-effect MEM:

$$\log T_{ij}=\boldsymbol{x}_{ij}^T\beta+\boldsymbol{z}_{ij}^{*T}\boldsymbol{u}_i+\epsilon_{ij}\ , \tag{2}$$

Here z_{ij} is the jth component of $Z_i^*(\rho)$ in (A4), $\boldsymbol{u}_i\sim N(0,\sigma_v^2 I_k)$ with $\sigma_v^2=\sigma_g^2+\sigma_c^2$. Here, I_k is the k-dimensional identity matrix such that $k=1$ for MZ_i and $k=2$ for DZ_i. For the inference on parameters in model (1) we use model (2), which is the form of single random-effect MEM by Ha et al. (2002). Thus, model (2) can be fitted using Ha et al.'s (2002) method, as we shall show in Section 4.

4. Estimation Procedure

Let F_{ij} be the censoring time corresponding to survival time T_{ij}. Let $Y_{ij}=\min(\log T_{ij},\log F_{ij})$ and $\delta_{ij}=I(T_{ij}\le F_{ij})$, where $I(\cdot)$ is the indicator function. Following Ha et al. (2002), the h-loglikelihood h for model (2) with censoring is defined by

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

where $\ell_{1ij} = \ell_{1ij}(\beta, \sigma_\epsilon^2, \rho; y_{ij}, \delta_{ij}|\boldsymbol{u}_i) = -\delta_{ij}\{\log(2\pi\sigma_\epsilon^2) + (m_{ij})^2\}/2 + (1-\delta_{ij})\log\{1-\Phi(m_{ij})\}$ is the logarithm of the conditional density function for Y_{ij} and δ_{ij} given $\boldsymbol{u}_i$,

$$\ell_{2i} = \ell_{2i}(\sigma_v^2; \boldsymbol{u}_i) = -\{\log\det(2\pi\sigma_v^2 I_k) + (\boldsymbol{u}_i^T \boldsymbol{u}_i/\sigma_v^2)\}/2$$

is the logarithm of the density function for $\boldsymbol{u}_i$. Here $m_{ij} = (y_{ij} - \mu_{ij})/\sigma_\epsilon$,

$$\mu_{ij} = E(\log T_{ij}|\boldsymbol{u}_i) = \boldsymbol{x}_{ij}^T\beta + \boldsymbol{z}_{ij}^{*T}\boldsymbol{u}_i$$

and Φ is the standard normal distribution function.

Note here that $E(Y_{ij}|\boldsymbol{u}_i) \neq \mu_{ij}$, Following Ha et al. (2002), the h-likelihood method uses the pseudo-responses y_{ij}^* such that

$$E(y_{ij}^*|\boldsymbol{u}_i) = \mu_{ij},$$

Here

$$\begin{aligned} y_{ij}^* &= E(\log T_{ij}|\, Y_{ij} = y_{ij}, \delta_{ij}, \boldsymbol{u}_i) \\ &= y_{ij}\delta_{ij} + A_{ij}(1-\delta_{ij}) \end{aligned}$$

where $A_{ij} = E(\log T_{ij}|\log T_{ij} > y_{ij}, \boldsymbol{u}_i) = \mu_{ij} + \sigma_\epsilon V(m_{ij})$ and $V(\cdot)$ is the hazard function for $N(0,1)$. Thus, for a moderate sample size the model (2) can be straightforwardly fitted using Ha et al.'s (2002) method. However, for the large data set such as the twin survival data of Section 2 the dimension of model matrix for random effects $\boldsymbol{u}_i$'s increases with q. In this case, it is difficult to apply directly the h-likelihood procedure of Ha et al. (2002), for example in solving the score equations of fixed- and random-effects.

We thus propose a simple and fast computation method for the large data using partition matrix as follows. Let μ be the $n \times 1$ vector with the ijth element μ_{ij},

$$\mu = X\beta + Z^*u,$$

where $X = (X_1^T, \cdots, X_q^T)^T$ is the $n \times p$ model matrix for the $p \times 1$ fixed effects β and $Z^* = \mathrm{BD}(Z_1^*, \cdots, Z_q^*)$ is $n \times q^*$ block diagonal matrix for $q^* \times 1$ random effects $\boldsymbol{u} = (\boldsymbol{u}_1^T, \cdots, \boldsymbol{u}_q^T)^T$. Here, $q^* = q_1 + 2q_2$, q_1 is the number of MZ twin pairs and q_2 is that of DZ twin pairs. Note that $q = q_1 + q_2$. Let $\boldsymbol{y}^* = (\boldsymbol{y}_1^{*T}, \cdots, \boldsymbol{y}_q^{*T})^T$ be the $n \times 1$ vector with ith vector $\boldsymbol{y}_i^* = (y_{i1}^*, y_{i2}^*)^T$. Assume that ρ is known. Given $\theta = (\sigma_\epsilon^2, \sigma_v^2)^T$ and $\boldsymbol{y}^*$, the maximum h-likelihood estimators of $\tau = (\beta^T, \boldsymbol{u}^T)^T$ becomes Henderson's (1975) mixed-model equations with pseudo-response variables $\boldsymbol{y}^*$ (Ha et al., 2002). Thus, we can easily show that substituting $X = (X_1^T, \cdots, X_q^T)^T$, $Z^* = \mathrm{BD}(Z_1^*, \cdots, Z_q^*)$ and $\boldsymbol{y}^* = (\boldsymbol{y}_1^{*T}, \cdots, \boldsymbol{y}_q^{*T})^T$ into the mixed-model equations reduces to the following score equations:

$$(\sum_i X_i^T X_i)\hat{\beta} = \sum_i X_i^T y_i^* - \sum_i X_i^T Z_i^* \hat{u}_i, \tag{3}$$

$$(Z_i^{*T} Z_i^* + \lambda I_k)\hat{u}_i = Z_i^{*T} y_i^* - Z_i^{*T} X_i \hat{\beta} \quad (i = 1, \cdots, q), \tag{4}$$

where $\lambda = \sigma_\epsilon^2/\sigma_v^2$. Following Lee and Nelder (1996) and Ha et al. (2002), the asymptotic variance for $\hat{\beta}$ is given by the upper left-hand coner, D^{11}, of the inverse of $D(h,\tau) = -\partial^2 h/\partial \tau^2$and for the large data it can be also expressed as follows:

$$D^{11} = \sigma_\epsilon^2 \left\{ \sum_i X_i^T W_i X_i - \sum_i (X_i^T W_i Z_i^*)(Z_i^{*T} W_i Z_i^* + \lambda I_k)^{-1}(Z_i^{*T} W_i X_i) \right\}^{-1},$$

where W_i is the ith component of $W = \mathrm{diag}(w_{ij})$, which is the $n \times n$ diagonal weight matrix with ijth element $w_{ij} = \delta_{ij} + (1-\delta_{ij})\xi(m_{ij})$ and $\xi(x) = V(x)\{V(x) - x\}$.

Let ℓ be a likelihood with nuisance effects ψ. Lee and Nelder (1996, 2001a) considered a function $p_\psi(\ell)$, defined by

$$p_\psi(\ell) = [\ell - \frac{1}{2} log \det \{D(\ell, \psi)/(2\pi)\}]|_{\psi = \hat{\psi}},$$

where $D(\ell, \psi) = -\partial^2 \ell/\partial \psi^2$ and $\hat{\psi}$ solves $\partial \ell/\partial \psi = 0$; $p_\psi(\ell)$ is an adjusted profile likelihood that eliminates the nuisance effects ψ from ℓ. For the estimation of the dispersion parameters $\psi = (\sigma_\epsilon^2, \sigma_v^2)^T$, Ha et al. (2002) used a restricted likelihood (or adjusted profile h-likelihood), $p_\tau(h) = [h - (1/2) log \det \{D(h,\tau)/(2\pi)\}]|_{\tau = \hat{\tau}}$, after eliminating fixed- and random-effects τ. However, the estimation of θ using $p_\tau(h)$ requires the inverse of $D(h,\tau) = -\partial^2 h/\partial \tau^2$, which could be computationally intensive in large samples. Recently, Noh and Lee (2004) showed that the resulting dispersion estimators from the restricted likelihoods $p_\tau(h)$ and $p_u(h) = [h - (1/2) log \det \{D(h,u)/(2\pi)\}]|_{u = \hat{u}}$, are asymptotically equivalent. In particular, the inversion of $D(h,u) = -\partial^2 h/\partial u^2$ in $p_u(h)$ is very simple because $D(h,u) = H_{22}/\sigma_\epsilon^2$ is a diagonal matrix with $H_{22} = Z^{*T} W Z^* + \Lambda$, where $\Lambda = \lambda I_{q^*}$ and I_{q^*} is the $q^* \times q^*$ identity matrix. In this paper, for the estimation of θ we use $p_u(h)$. This leads to the ML (maximum likelihood) type estimators for σ_ϵ^2 and σ_v^2, given by

$$\hat{\sigma}_\epsilon^2 = \sum_{ij} (y_{ij}^* - \hat{\mu}_{ij})^2 / \{n_1 - (q^* - \gamma_1)\} \quad \text{and} \quad \hat{\sigma}_v^2 = \sum_i \hat{u}_i^T \hat{u}_i / (q^* - \gamma_2) \tag{5}$$

where $n_1 = \sum_{ij} w_{ij}$, $\gamma_1 = \sigma_\epsilon^2 \, \mathrm{trace} \left\{ H_{22}^{-1} (\partial H_{22}/\partial \sigma_\epsilon^2) \right\}$ and $\gamma_2 = -\sigma_v^2 \, \mathrm{trace} \left\{ H_{22}^{-1} (\partial H_{22}/\partial \sigma_v^2) \right\}$. The formulation for the $\partial H_{22}/\partial \sigma_\epsilon^2$ and $\partial H_{22}/\partial \sigma_v^2$ terms are given in Appendix II of Ha et al. (2002) and the trace terms in γ_1 and γ_2 are easily calculated using the partition matrix. Note that since we cannot observe all the y_{ij}^*'s due to the censoring, we substitute estimates, say $\hat{y}_{ij}^*$,

for them in each iteration.

The fitting algorithm is summarized as follows:

(Step 1) Given ρ (and hence $L_i(\rho)$), estimate τ and θ using (3), (4) and (5).

(Step 2) Given τ and θ estimate ρ by maximizing $p_u(h)$.

(Step 3) Iterate **(Step 1)** and **(Step 2)** until convergence is achieved.

After convergence has occurred, we compute the estimates of σ_g^2 and σ_c^2 from (A5) and those of $var(\hat{\beta})$ from D^{11}, respectively.

5. Application

To illustrate the proposed method, we analyze the twin survival data of the old cohort in Section 2. Firstly, we analyze separately MZ and DZ twins in males (females). For this we use the model (1) without random-environment effects c_{ij} because of an identifiable problem on estimation of dispersion parameters. The fitted results are given in Table 3.

<Table 3> Separate analyses using the genetic models for the old cohort

Data	Model	$\hat{\beta}_0$(SE)	$\widehat{\sigma_g^2}$	$\widehat{\sigma_\epsilon^2}$	$\widehat{h_g^2}$	$-2p_u(h)$
Males						
MZ	E	4.356(0.0026)	—	0.0206	—	-1829.5
	GE	4.355(0.0030)	0.0069	0.0134	0.34	-2027.2
DZ	E	4.345(0.0020)	—	0.0215	—	-3048.7
	GE	4.344(0.0021)	0.0054	0.0160	0.25	-3118.4
Females						
MZ	E	4.408(0.0024)	—	0.0198	—	-1434.7
	GE	4.406(0.0026)	0.0051	0.0144	0.26	-1565.6
DZ	E	4.404(0.0017)	—	0.0204	—	-2764.0
	GE	4.403(0.0018)	0.0038	0.0164	0.19	-2818.1

β_0, intercept; SE, the corresponding standard error; E, $\log T = \beta_0 + \epsilon$ with $\sigma_g^2 = 0$; GE, $\log T = \beta_0 + g + \epsilon$ with $\sigma_g^2 > 0$; $h_g^2 = \sigma_g^2/(\sigma_g^2 + \sigma_\epsilon^2)$.

For testing the need for a random component (i.e., $\sigma_g^2 = 0$), we use the deviance ($-2p_u(h)$ in Table 3) based upon the restricted likelihood $p_u(h)$ because $p_\tau(h)$ and $p_u(h)$ give asymptotically equivalent dispersion estimates. Note that $p_u(h)$ is the first-order Laplace approximation to marginal likelihood m and that it would be natural to use $p_u(h)$ when m is numerically hard to obtain (Lee and Nelder, 2001a). Because such a hypothesis ($H_0 : \sigma_g^2 = 0$) is on the boundary of the parameter space the critical value is $\chi_{2\alpha}^2$ for a size α test. This value results from the fact that the asymptotic distribution of likelihood ratio test is a 50:50 mixture

of χ_0^2 and χ_1^2 distributions (Self and Liang, 1987): for applications to random-effect models see Vu and Knuiman (2002) and Ha and Lee (2005a). For example, for male MZ twins the deviance difference between the E and GE models is 197.7, which is significant at a 5% level ($\chi_{1,0.10}^2 = 2.71$ due to the half χ^2-distribution), indicating that the random-genetic effects are necessary, i.e., $\sigma_g^2 > 0$. Furthermore, we see that the random-genetic effects are all necessary in each data set of Table 3. As expected, from the value of $\hat{\beta}_0$ in the GE model we observe that the male MZ[DZ] twin tends to have shorter life span than in female MZ[DZ] twin, respectively; for example, the estimated mean life span is $exp(4.355) = 77.9$(ages) for male MZ and $exp(4.406) = 81.9$(ages) for female MZ. On the other hand, from the estimated heritability $\widehat{h_g^2}$ of four GE models in Table 3 we also see that in both sexes the higher heritability in MZ over DZ is interpreted as evidence that MZ twins share more longevity-related genetic material than DZ twins: see also Yashin and Iachine (1995).

Table 4 shows the results of fitting the model (1) with a single fixed covariate x_{ij}(=1 for MZ_i and =0 for DZ_i) when data for MZ and DZ twins are combined according to sexes. For testing the need for a random component (i.e., $\sigma_g^2 = 0$ or $\sigma_c^2 = 0$), we again use the deviance ($-2p_u(h)$) as in Table 3. We first analyze the male data set. The deviance difference between GE and GCE is 0.00, which is not significant at a 5% level ($\chi_{1,0.10}^2 = 2.71$), indicating the absence of the random-environmental effects (i.e., $\sigma_c^2 = 0$). The deviance difference between CE and GCE is 37.7, indicating that the random-genetic effects are necessary, i.e., $\sigma_g^2 > 0$. In addition, the deviance difference between E and GE is 264.6, indicating that the random-genetic effects are indeed necessary with or without random-environmental effects. Again, the results obtained from the female data set are similar to those evident in the male data set.

<Table 4> Combined (MZ and DZ) analyses using the genetic models for the old cohort

Data	Model	$\hat{\beta}_0$(SE)	$\hat{\beta}_1$(SE)	$\widehat{\sigma_g^2}$	$\widehat{\sigma_c^2}$	$\widehat{\sigma_\epsilon^2}$	$\widehat{h_g^2}$	$-2p_u(h)$	AIC
Males	E	4.345(0.0020)	0.012(0.0033)	—	—	0.0212	—	-4876.7	262.6
	CE	4.344(0.0021)	0.012(0.0036)	—	0.0046	0.0165	—	-5103.6	37.7
	GE	4.343(0.0021)	0.012(0.0036)	0.0066	—	0.0144	0.32	-5141.3	0
	GCE	4.343(0.0021)	0.012(0.0036)	0.0066	0.0000	0.0144	0.32	-5141.3	2.0
Females	E	4.404(0.0017)	0.004(0.0032)	—	—	0.0202	—	-4197.9	181.5
	CE	4.403(0.0018)	0.005(0.0032)	—	0.0033	0.0167	—	-4370.3	11.1
	GE	4.402(0.0018)	0.005(0.0032)	0.0047	—	0.0153	0.24	-4381.4	0
	GCE	4.402(0.0018)	0.005(0.0032)	0.0047	0.0000	0.0153	0.24	-4381.4	2.0

β_0, intercept; β_1, MZ effect; SE, the corresponding standard error; E, MEM with $\sigma_g^2 = \sigma_c^2 = 0$; CE, MEM with $\sigma_g^2 = 0, \sigma_c^2 > 0$; GE, MEM with $\sigma_g^2 > 0, \sigma_c^2 = 0$; GCE, MEM with $\sigma_g^2 > 0, \sigma_c^2 > 0$; $h_g^2 = \sigma_g^2/(\sigma_g^2 + \sigma_c^2 + \sigma_\epsilon^2)$; AIC, the difference of AIC.

To select a model among nested models a testing procedure such as the above can be used.

However, for a model selection among non-nested models such as CE and GE we can also consider the following Akaike information criterion (AIC)

$$\text{AIC} = -2p_u(h) + 2d, \tag{6}$$

where d is the number of fixed and dispersion parameters, not the number of random effects.

Since $p_\tau(h)$ and $p_u(h)$ give asymptotically equivalent dispersion estimators, the AIC of (6) is an extension of the AIC in SAS PROC MIXED (Wolfinger, 1993) based upon restricted likelihood for selecting a specific covariance structure in mixed linear models: for the use of AIC based on $p_\tau(h)$ see Ha and Lee (2005a) and Noh et al. (2005). We may select the model which has the smallest AIC value. For ease of comparison and ranking of candidate models, we set the smallest value to be zero. From Table 4, for the male data the AIC chooses the GE as the best model, with $\widehat{h_g^2} = 32\%$. Again, the GE model fits the female data best, with $\widehat{h_g^2} = 24\%$. Previously, Yashin et al. (1999) analyzed a Swedish twin survival data set, similar to the old cohort in Table 2 using various frailty models. Notice here that the current data set is more complete due to more recent follow-up study. By using an AIC Yashin et al. (1999) also chose a genetic frailty model, corresponding to the GE model, as a final model. However, our estimates on h_g^2 are different from those obtained from the final model by Yashin et al. (1999), which give 58% and 39% for males and females, respectively. However, in frailty models the estimates of dispersion parameters can be sensitive against misspecification of random-effect distribution (Agresti et al. 2004; Ha and Lee, 2005b), and thus the values of heritability estimated from the frailty models may not robust.

On the other hand, from the estimation of β_1 in both GE models of Table 4 we observe an interesting finding that in male twins the MZ ($\widehat{\beta_1} = 0.012$ with SE = 0.0036) tends to have significantly longer life span than in the DZ, but that in female twins this is no longer significant ($\widehat{\beta_1} = 0.005$ with SE =0.0032).

6. Simulation Study

We present a numerical study, based upon 200 replications of simulated data, to evaluate the performance of the proposed procedure in Section 4. Using the structures of the first data set (male twins data with a single covariate) in Table 4, the data are generated from a GCE model (2). That is, the random effects $\boldsymbol{u}_i$, $i = 1, \cdots, 4598$, are generated from $N(0, \sigma_v^2 I_k)$. Note here that we have $\boldsymbol{u}_i\,(= u_{i1} = u_{i2})$ only for the $\text{MZ}_i\,(k=1)$ and $\boldsymbol{u}_i = (u_{i1}, u_{i2})^T$ for the $DZ_i(k=2)$. Given $\boldsymbol{u}_i$, the survival times T_{ij}, $j = 1, 2$ are generated from $N(\mu_{ij}, \sigma_\epsilon^2)$ with $\mu_{ij} = \beta_0 + \beta_1 x_{ij} + \boldsymbol{z}_{ij}^{*T}\boldsymbol{u}_i$. We set the covariate x_{ij} to be 1 for MZ_i $(i = 1, \cdots, 1646)$ and to be 0

for DZ_i $(i = 1647, \cdots, 4598)$. For the true parameters we use the estimates from GCE model in the third data set of Table 4; $\rho = 0.5$ (i.e., $\sigma_c^2 = 0$), $\beta_0 = 4.343$, $\beta_1 = 0.012$, $\sigma_v^2 = 0.0066$, $\sigma_\epsilon^2 = 0.0144$. Under this setting, we also set $\rho = 0.6$ to show another separation of σ_g^2 and σ_c^2 ; from (A.5) this gives $\sigma_g^2 = 0.0053$, $\sigma_c^2 = 0.0013$ and hence $h_g^2 = 0.0053/(0.0053 + 0.0013 + 0.0144) = 0.25$. The corresponding censoring times F_{ij}/θ_F are generated from standard exponential distribution with mean 1. Here, the values of parameters θ_F are empirically determined to achieve approximately the right censoring rate (around 20%). From 200 replications of simulated data we compute the mean, the standard deviation and the mean of the estimated standard errors for $\hat{\beta}_j$, $j = 0, 1$. The standard deviation (SD) for $\hat{\beta}_j$, $j = 0, 1$ is defined by SD $= \sum_i \{(\hat{\beta}_j^{(i)} - \overline{\beta}_j)^2/199\}^{1/2}$, where $\hat{\beta}_j^{(i)}$ is the estimate of β_j in the ith replication and $\overline{\beta}_j = \sum_i \hat{\beta}_j^{(i)}/200$ is the mean of the values of $\hat{\beta}_j^{(i)}$. The estimate of standard-error for $\hat{\beta}_j$ is obtained from D^{11}. For the dispersion parameters the mean and standard deviation for $\hat{\rho}$, $\hat{\sigma}_g^2$, $\hat{\sigma}_c^2$, $\hat{\sigma}_\epsilon^2$ and $\hat{h}_g^2$ are also given. For the computation we used SAS/IML. The results are summarized in Table 5.

<Table 5> Simulation results on estimation of parameters in the genetic model.

	$\hat{\rho}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}_g^2$	$\hat{\sigma}_c^2$	$\hat{\sigma}_\epsilon^2$	$\hat{h}_g^2$
True	0.50	4.343	0.012	0.0066	0	0.0144	0.32
Mean	0.505*	4.343	0.012	0.0065	6.6×10^{-5}	0.0145	0.31
SD		0.0023	0.0038	0.0006	6.4×10^{-6}	0.0005	0.026
	(0.0096*)	(0.0022)	(0.0038)				
True	0.6	4.343	0.012	0.0053	0.0013	0.0144	0.25
Mean	0.602	4.343	0.012	0.0053	0.0013	0.0144	0.25
SD	0.0784	0.0022	0.0037	0.0013	0.0009	0.0006	0.063
		(0.0022)	(0.0038)				

The simulation is conducted with 200 replications under the structure of the first data set ($n = 4598$ male twin pairs with about 20% censoring) in Table 4;
Mean and SD indicate the mean and standard deviation for estimates of each parameter;
* : ML estimate and SE of ρ from fitting the classical Tobit model for the sample of 200 simulated estimates of ρ under$\rho = 0.5$; ( ): the mean of estimated SE's; $h_g^2 = \sigma_g^2/(\sigma_g^2 + \sigma_c^2 + \sigma_\epsilon^2)$.

When $\rho = 0.5$ (i.e., $\sigma_c^2 = 0$) it is on boundary of the parameter space $(0.5 \le \rho \le 1.0)$ and the asymptotic distribution of the estimates is a 50:50 mixture between a point mass at 0.5 and a normal random variable on the axis large than 0.5 (Vu and Knuiman, 2002). In the current

simulation we also confirm that the observed proportion of 0.5 estimates out of 200 simulated estimates of ρ was approximately equal to 50% because of 103/200=0.515. In case that $\rho = 0.5$, for the dispersion parameters (σ_g^2, σ_c^2 and h_g^2) we report the estimation results based on a ML estimate of ρ using the sample of 200 simulated estimates of ρ. In order to obtain the ML estimate of ρ, we use the classical Tobit model (i.e., a regression model with left-censoring; Tobin, 1958), which is fitted via SAS PROC LIFEREG. From Table 5 we make the following observations. Overall, the interested h-likelihood estimates $\hat{\beta}_j$, $\hat{\sigma}_g^2$, $\hat{\sigma}_c^2$, $\hat{\sigma}_\epsilon^2$ and $\hat{h}_g^2$ work well. In Table 5 SD is the estimate of the true $\{\mathrm{var}(\hat{\beta}_j)\}^{1/2}$ and SE is the average of standard-error estimates for $\hat{\beta}_j$. Our standard-error estimates work well as judged by the very good agreement between SE and SD. These results suggest that the proposed method is indeed reasonable.

The proposed method of this paper is parametric. However, with the use of MEMs we see from Table 4 that the estimates for fixed effects β are insensitive to the choice of dispersion models E, CE, GE or GCE. Furthermore, Ha et al. (2002) showed by simulation study that the MEMs give insensitive inferences against the misspecification of the distribution of random effects if the censoring rate is not too high. Moreover, each data set of the old cohort in Table 2 is overall in the range of a low censoring rate as 20%~30%.

Appendix: *Derivation of model (2)*

From $v_{ij} = g_{ij} + c_{i0}$ for $j = 1, 2$, the model (1) can be expressed as a simple matrix form

$$\log \boldsymbol{T}_i = X_i\beta + Z_i\boldsymbol{v}_i + \boldsymbol{\epsilon}_i \quad , \tag{A1}$$

where $\boldsymbol{T}_i = (T_{i1}, T_{i2})^T$, $X_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2})^T$ is the $2 \times p$ model matrix of β, Z_i is the model matrix of $\boldsymbol{v}_i$, $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \epsilon_{i2})^T \sim N(0, \sigma_\epsilon^2 I_2)$ and I_2 is the 2×2 identity matrix. Note here that for the MZ_i $Z_i = (1,1)^T$ and $\boldsymbol{v}_i\,(= v_{i1} = v_{i2}) \sim N(0, \sigma_v^2)$, but that for the DZ_i $Z_i = I_2$ and $\boldsymbol{v}_i = (v_{i1}, v_{i2})^T \sim N(0, \sigma_v^2 \Sigma_i)$ with a compound symmetric structure $\Sigma_i = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

Here

$$\sigma_v^2 = \sigma_g^2 + \sigma_c^2 \quad , \tag{A2}$$

$$\rho = \mathrm{corr}\,(v_{i1}, v_{i2}) = \frac{0.5\sigma_g^2 + \sigma_c^2}{\sigma_g^2 + \sigma_c^2} \quad , \tag{A3}$$

where $\rho \in [0.5, 1.0]$. The use of ρ leads to useful results. From (A3) we have that σ_g^2 is very larger than σ_c^2 (i.e., $\sigma_g^2 \gg \sigma_c^2$) as ρ goes to 0.5, but that $\sigma_g^2 \ll \sigma_c^2$ as ρ goes to 1.0. In particular, the model (A1) reduces to a model (1) without random-environment effects c_{ij} if $\rho = 0.5$ (i.e., $\sigma_c^2 = 0$), while it does that without random-genetic effects g_{ij} if $\rho = 1.0$ (i.e.,

$\sigma_g^2 = 0$). Following Lee and Nelder (2001b), the random effects v_i for DZ_i are assumed to have the form $L_i(\rho)u_i$, where $u_i \sim N(0, \sigma_v^2 I_2)$. For the DZ_i, from Cholesky decomposition we have a lower triangular matrix L_i such that $\Sigma_i = L_i L_i^T$. Here, we choose

$$L_i = L_i(\rho) = \begin{pmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{pmatrix},$$

and so the random effects $v_i = L_i u_i \sim N(0, \sigma_v^2 L_i L_i^T)$. Thus, the model (A1) can be written as

$$\log T_i = X_i\beta + Z_i^* u_i + \epsilon_i, \tag{A4}$$

where $u_i \sim N(0, \sigma_v^2 I_k)$, and $Z_i^* = (1,1)^T$ and $I_k = 1$ for the MZ_i, and $Z_i^* = L_i(\rho)$ and $I_k = I_2$ for the DZ_i. Note that from (A2) and (A3) we obtain σ_g^2 and σ_c^2 as follows:

$$\sigma_g^2 = \sigma_v^2 - \sigma_c^2 \quad \text{and} \quad \sigma_c^2 = 2(\rho - 0.5)\sigma_v^2. \tag{A5}$$

Then the jth element of model (A4) becomes the model (2).

References

[1] Agresti, A., Caffo, B. and Ohman-Strickland, P. (2004). Example in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics and Data Analysis*, 47, 639-653.

[2] Ha, I. D. and Lee, Y. (2005a). Multilevel mixed linear models for survival data. *Lifetime Data Analysis*, 11, 131-142.

[3] Ha, I. D. and Lee, Y. (2005b). Comparison of hierarchical likelihood versus orthodox best linear unbiased predictor approaches for frailty models. *Biometrika*, 92, 717-723.

[4] 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.

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

[6] Hougaard, P. (2000). *Analysis of multivariate survival data*. Springer-Verlag, New York.

[7] Lambert, P., Collett, D., Kimber, A, and Johnson, R. (2004). Parametric accelerated failure time models with random effects and an application to kidney transplant survival. *Statistics in Medicine*. 23, 3177-3192.

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

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

[10] Lee, Y. and Nelder, J. A. (2001b). Modelling and analysing correlated non-normal data. *Statistical Modelling*, 1, 3-16.

[11] Neal, M. C. and Cardon, L. R. (1992). *Methodology for genetic studies of twin and families.*

Kluwer Academic: Dordrecht.

[12] Noh, M. and Lee, Y. (2004). REML estimation for binary data in GLMMs. A manuscript submitted for publication.

[13] Noh, M., Ha, I. D. and Lee, Y. (2005). Dispersion frailty models and HGLMs. *Statistics in Medicine*, 24, in press.

[14] Pawitan, Y., Reilly, M., Nilsson, E, Cnattingius, S. and Lichtenstein, P. (2004). Estimation of genetic and environmental factors for binary traits using family data. *Statistics in Medicine,* 23, 449-465.

[15] Self, S. G. and Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82, 605-610.

[16] Sham, P. C. (1998). *Statistics in Human Genetics*. Arnold: London.

[17] Tobin, J. (1958). Estimation of relationship for limited dependent variables. *Econometrica*, 26, 24-36.

[18] Vu, H. T. V. and Knuiman, M. W. (2002). A hybrid ML-EM algorithm for calculation of maximum likelihood estimates in semiparametrics shared frailty models. *Computational Statistics and Data Analysis*, 40, 173-187.

[19] Wolfinger, R. D. (1999). Fitting nonlinear mixed models with the new NLMIXED procedure. *Proceedings of the 99 Joint Statistical Meetings*, 287.

[20] Yashine, A. I., Iachine, I. A. (1995). How long can humans live? Lower bound for biological limit of human longevity calculated from Danish twin data using correlated frailty model. *Mechanisms of Ageing and Development*, 80, 147-169.

[21] Yashine, A. I., Iachine, I. A. and Harris, J. R. (1999). Half of the variation in susceptibility to mortality is genetic: findings from Swedish twin survival data. *Behavior Genetics*, 29, 11-19.

[Received August 2005, Accepted November 2005]