Journal of the Korean Data & Information Science Society
2009, **20**(1), 203–209

한국데이터정보과학회지

On prediction of random effects in log-normal frailty models[†]

IL Do Ha[1] · Geon-Ho Cho[2]

[12]Department of Asset Management, Daegu Haany University

Received 12 December 2008, revised 20 January 2009, accepted 23 January 2009

Abstract

Frailty models are useful for the analysis of correlated and/or heterogeneous survival data. However, the inferences of fixed parameters, rather than random effects, have been mainly studied. The prediction (or estimation) of random effects is also practically useful to investigate the heterogeneity of the hospital or patient effects. In this paper we propose how to extend the prediction method for random effects in HGLMs (hierarchical generalized linear models) to log-normal semiparametric frailty models with nonparametric baseline hazard. The proposed method is demonstrated by a simulation study.

Keywords: Frailty models, hierarchical generalized linear models, hierarchical likelihood, prediction interval, random effects.

1. Introduction

Frailty models, extensions of Cox's (1972) proportional hazard (PH) models, are very useful for the analysis of survival data with correlated and/or heterogeneous structures (Hougaard, 2000; Duchateau and Janssen, 2008). Here, the frailty means an unobserved random effect in the hazard models. However, the inferences of fixed parameters, rather than random effects, have been mainly studied. The prediction (or estimation) of random effects is also practically useful to investigate the heterogeneity of cluster effects such as hospital or patient effects (Vaida and Xu, 2000).

For the inference of frailty models an intractable integration is often required (Ha, Lee and Song, 2001). Thus, the hierarchical likelihood (h-likelihood; Lee and Nelder, 1996) can be widely used because it provides a statistically efficient procedure in various random-effect models such as HGLMs and frailty models (Lee and Nelder, 2001; Lee, Nelder and Pawitan, 2006; Ha and Lee, 2005).

[†] This work was partially supported by the Korea Research Foundation Grant funded by the Korean Government(KRF-2008-521-C00057).

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

[2] Professor, Department of Asset Management, Daegu Haany University, Gyeongsan 712-715, Korea.

In this paper we propose a h-likelihood method to construct prediction interval in log-normal semiparametric frailty models. For this we need to find a proper standard error of random effects. The empirical Bayes method based on the conditional distribution of random effects has been used, but it underestimates the standard error (Vaida and Xu, 2000). We show that the h-likelihood procedure accounts for the inflation of standard error of random effects caused by uncertainty in the estimation of fixed parameters (i.e. regression and dispersion parameters). In fact, this is an extension of the prediction method for random effects in Poisson HGLMs (Lee and Nelder, 1996; Ha, 2008a) to log-normal frailty models with nonparametric baseline hazard. The inference on prediction interval is also important for relative-risk analysis in disease mapping problem (Ainsworth and Dean, 2006) or multi-center clinical trial (Vaida and Xu, 2000). Simulation study shows that for the prediction intervals the h-likelihood method maintains well the required level.

The paper is organized as follows. In Section 2 we briefly describe the frailty models. In Section 3 we outline the h-likelihood estimation procedure on the estimation of fixed parameters as well as random effects and then show how to obtain the prediction intervals. Finally, simulation study is given in Section 4.

2. Frailty Models

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

$$y_{ij} = \min(T_{ij}, C_{ij}) \text{ and } \delta_{ij} = I(T_{ij} \leq C_{ij}),$$

where $I(\cdot)$ is the indicator function. Denote by v_i the unobserved frailty (or random effect) for the i th cluster. Given v_i, the conditional hazard function of T_{ij} takes the form

$$\lambda_{ij}(t|v_i) = \lambda_0(t) \exp(x_{ij}^T \beta + v_i), \tag{2.1}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function, β is a $p \times 1$ vector of unknown regression parameters corresponding to fixed covariates $x_{ij} = (x_{ij1}, \ldots, x_{ijp})^{\mathrm{T}}$. Here, the term $x_{ij}^T \beta$ in (2.1) does not include an intercept term because of identifiable purposes. The distribution of frailties v_i 's is commonly assumed to follow a normal distribution with mean $E(v_i) = 0$ and var $(v_i) = \alpha$ (McGilchrist, 1993; Ha et al., 2001). In particular, the normal assumption for v_i is very useful for modelling multi-component (Ha, Lee and MacKenzie, 2007) or correlated frailties (Vaida and Xu, 2000). For the v_i other frailty distribution such as log-gamma can be assumed (Hougaard, 2000). Note that if v_i is normal or log-gamma, $\exp(v_i)$ becomes log-normal or gamma, respectively, and the corresponding model is called log-normal or gamma frailty model: see also Hougaard (2000). Here the variance α is called frailty or dispersion parameter. Notice also that the model (2.1) becomes Cox's PH model if $\alpha = 0$ (i.e. $v_i \equiv 0$).

Since the functional form of $\lambda_0(t)$ is unknown, following Breslow (1972) and Ha et al., (2001) we consider the baseline cumulative hazard function $\Lambda_0(t)$ to be a step function with jumps at the r distinct observed death times,

$$\Lambda_0(t) = \sum_{k: y_{(k)} \leq t} \lambda_{0k},$$

where $y_{(k)}$ is the k th ($k = 1, \cdots, r$) smallest distinct death time among the y_{ij}'s, and $\lambda_{0k} = \lambda_0(y_{(k)})$.

3. H-likelihood Approach

In this section we present the h-likelihood method for the inference of frailty models. Firstly, we outline the corresponding estimation procedure. Then we show how to find the standard error of random-effect estimator and to construct the prediction interval for random effects.

3.1. Estimation procedure

Following Ha et al. (2001), the h-likelihood for the log-normal frailty model (2.1) is defined by

$$h = h(\beta, \lambda_0, \alpha) = \sum_{ij} \ell_{1ij} + \sum_{i} \ell_{2i},$$

where

$$\begin{aligned}\sum_{ij} \ell_{1ij} &= \sum_{ij} \delta_{ij} \left\{\log \lambda_0(y_{ij}) + \eta_{ij}\right\} - \sum_{ij} \Lambda_0(y_{ij}) \exp(\eta_{ij}) \\ &= \sum_{k} d_{(k)} \log \lambda_{0k} + \sum_{ij} \delta_{ij}\eta_{ij} - \sum_{k} \lambda_{0k} \left\{ \sum_{(i,j)\in R_{(k)}} \exp(\eta_{ij}) \right\},\end{aligned}$$

$\ell_{1ij} = \ell_{1ij}(\beta, \lambda_0; y_{ij}, \delta_{ij}|v_i)$ is the logarithm of the conditional density function for y_{ij} and δ_{ij} given v_i ,

$$\ell_{2i} = \ell_{2i}(\alpha; v_i) = -\frac{1}{2}\log(2\pi\alpha) - \frac{1}{2\alpha} v_i^2$$

is the logarithm of the density function for v_i with parameter α , $\lambda_0 = (\lambda_{01}, \ldots, \lambda_{0r})^T$, $\eta_{ij} = x_{ij}^T\beta + v_i$, $d_{(k)}$ is the number of deaths at $y_{(k)}$ and $R_{(k)} = R(y_{(k)}) = \left\{(i,j) : y_{ij} \geq y_{(k)}\right\}$ is the risk set at $y_{(k)}$.

Since the dimension of λ_0 increases with sample size n , for the estimation of (β, v) with $v = (v_1, \ldots, v_q)^T$ Ha et al. (2001) and Ha and Lee (2003) proposed to the use of the profile h-likelihood h^* with λ_0 eliminated:

$$h^* = h|_{\lambda_0 = \hat{\lambda}_0} = \sum_{ij} \delta_{ij}\eta_{ij} - \sum_{k} d_k \log \left\{ \sum_{(i,j)\in R_{(k)}} \exp(\eta_{ij}) \right\} + \sum_{i} \ell_{2i}, \tag{3.1}$$

where

$$\hat{\lambda}_{0k} = \frac{d_{(k)}}{\sum_{(i,j)\in R(y_{(k)})} \exp(x_{ij}^T\beta + v_i)}$$

are solutions of the estimating equations, $\partial h/\partial\lambda_{0k} = 0$, for $k = 1, \ldots, r$. Note that for log-normal and gamma frailty models, h^* becomes the kernel of the penalized partial likelihood (Ripatti and Palmgren, 2000).

With h^* in (3.1) we estimate fixed parameters (β, α) and random effects v as follows. Ha et al. (2001) further showed that given α the estimation of $\tau = (\beta, v)^T$ is obtained by solving

$$\partial h^*/\partial \tau = (\partial h/\partial \tau)|_{\lambda_0=\hat{\lambda}_0} = 0. \tag{3.2}$$

Next, for the estimation of the frailty parameter α we use Lee and Nelder's (2001) adjusted profile h-likelihood, defined by

$$p_\tau(h^*) = [h^* - \frac{1}{2}\log\det\{H(h^*,\tau)/(2\pi)\}]|_{\tau=\hat{\tau}},$$

where $H(h^*,\tau) = -\partial^2 h^*/\partial\tau^2$ and $\hat{\tau}$ solves $\partial h^*/\partial\tau = 0$ in (3.2). The h-likelihood estimator, an extension of restricted maximum likelihood (REML) estimator, for α is obtained by solving iteratively

$$\partial p_\tau(h^*)/\partial\alpha = 0. \tag{3.3}$$

In summary, the estimates $(\hat{\tau}, \hat{\alpha})$ are obtained by the alternation between the two estimating equations (3.2) and (3.3) until convergence is achieved.

3.2. Prediction intervals for random effects

Lee and Nelder (1996) noted that in HGLMs the location parameters (β, v) and dispersion parameter α are orthogonal. In particular, Poisson HGLMs can be expressed to the frailty models (Ha and Lee, 2003, 2005). Thus, with h^* in (3.1) we need to consider only the variance inflation caused by estimating β . Note that the asymptotic covariance matrix of $\hat{\beta}$ and $\hat{v} - v$ is the inverse of Hessian matrix H without nuisance parameters λ_0 (Ha et al., 2001 and Ha and Lee, 2003), given by

$$H(\beta, v) = -\begin{pmatrix} \partial^2 h^*/\partial\beta^2 & \partial^2 h^*/\partial\beta\partial v \\ \partial^2 h^*/\partial v\partial\beta & \partial^2 h^*/\partial v^2 \end{pmatrix} = \begin{pmatrix} X^T W^* X & X^T W^* Z \\ Z^T W^* X & Z^T W^* Z + R \end{pmatrix} \tag{3.4}$$

where X is the $n \times p$ matrix whose i th row vector is x_{ij}^T , Z is the $n \times q$ group indicator matrix whose i th row vector is z_{ij}^T , W^* is the $n \times n$ symmetric matrix given in Appendix 2 of Ha and Lee (2003) and $R = \text{diag}\left\{-\partial^2 \ell_{2i}/\partial v_i^2\right\}$ is the $q \times q$ diagonal matrix. The upper left-hand corner of H^{-1} in (3.4), provides an variance of $\hat{\beta}$, given by

$$\text{var}(\hat{\beta}) = (X^T V^{-1} X)^{-1} \text{ with } V = W^{*-1} + ZR^{-1}Z^T,$$

and the bottom right-hand corner of H^{-1} also gives an variance of $\hat{v} - v$, given by

$$\text{var}(\hat{v} - v) = \left\{(Z^T W^* Z + R) - (Z^T W^* X)(X^T W^* X)^{-1}(X^T W^* Z)\right\}^{-1}. \tag{3.5}$$

Thus we construct the prediction interval for random effects as follows. For the 95% prediction interval for v_i $(i = 1, \ldots, q)$ under asymptotic normality of the estimators, we have that

$$\hat{v_i} \pm 1.96 \times \text{SE}(\hat{v_i}), \tag{3.6}$$

where $\mathrm{SE}(\hat{v}_i) = \sqrt{\mathrm{var}(\hat{v}_i - v_i)}$ is the estimated standard error obtained from the inverse of information matrix H in (3.4). In Poisson HGLMs, H^{-1} gives a proper standard-error estimate for estimators of random effects (Ha, 2008a). For the SE Vaida and Xu (2000) used the empirical Bayes method, based on distribution of $v|(y,\delta)$. Under the normality of $\hat{v} - v$, $\mathrm{var}\{v|(y,\delta)\}$ can be estimated by $-(\partial^2 h^*/\partial v^2)^{-1}|_{\alpha=\hat{\alpha}}$. When α is known, $\mathrm{var}\{v|(y,\delta)\}$ is the only sensible prediction variance for v . However, α is unknown it underestimates variance of random-effect estimates, ignoring variability due to estimates of α (Vaida and Xu, 2000; Ha, 2008a).

4. Simulation Study

Simulated studies, based upon 500 replications of simulated data, are presented to evaluate the performance of the proposed method. We generate data from the frailty model (2.1) assuming the exponential baseline hazard $\lambda_0(t) = 1$, a regression parameter $\beta = 1$ and the variance $\alpha \equiv \sigma^2 = 0.25$. Here, we set a single covariate x_{ij} to be 0 for the first $q/2$ individuals (control group), and x_{ij} to be 1 for the remaining $q/2$ individuals (treatment group). We also set the sample size $n = \sum_{i=1}^{q} n_i = 100, 400, 800$ with $q = 25, 100, 200$ and $n_i = 4$. The corresponding censoring times C_{ij} are generated from an exponential distribution with parameter empirically determined to achieve approximately the right censoring rate, around 20%.

For the 500 replications we computed the mean, standard deviation (SD), the mean of the estimated standard error (SE) for $\hat{\beta}$. The SE is obtained from H^{-1} in (3.4). For the frailty parameter σ^2 the corresponding mean and SD are also given. For the model fitting and computation we used SAS/IML.

The results of parameter estimates are summarized in Table 4.1. As expected by Section 3.1, the h-likelihood estimates for fixed parameters (β, σ^2) work well as sample size n increases. In Table 4.1 SD is the estimate of the true $\{\mathrm{var}(\hat{\beta})\}^{1/2}$ and SE is the average of standard-error estimate for $\hat{\beta}$. Our standard-error estimate also performs well as judged by the very good agreement between SE and SD: see also the simulation results by Ha and Lee (2003, 2005). Note: The simulation is conducted with 500 replications for the log-normal

Table 4.1 Simulations results for the estimation of parameters in the frailty models

	$\hat{\beta}$				$\hat{\sigma^2}$		
n	Bias	SD	SE	MSE	Bias	SD	MSE
100	0.105	0.321	0.328	0.113	0.068	0.242	0.063
400	0.007	0.165	0.160	0.027	0.015	0.108	0.012
800	0.001	0.106	0.105	0.011	0.003	0.071	0.005

frailty models assuming the true regression parameter $\beta = 1$ and frailty variance $\sigma^2 = 0.25$, with three sample sizes $n = \sum_{i=1}^{q} n_i = 100, 400, 800$ (i.e. $q = 25, 100, 200$ and $n_i = 4$). SD, standard deviation of estimates over 500 simulations; SE, average of 500 estimated standard errors.

Furthermore, we are very interested in prediction intervals of random effects. From (3.6) we computed their 95% prediction intervals. That is, we calculated the coverage probabilities of the prediction intervals. The q samples $(v_1^{(k)}, \ldots, v_q^{(k)})$ which are generated for each replication $(k = 1, \ldots, 500)$ are treated as true random effects. We obtain the corresponding

coverage percentage from the 500 prediction intervals, so that for each method the number of the resulting coverage percentages is q . The results are plotted in Figure 4.1 This indicates that the h-likelihood method provides a proper prediction intervals as sample size n (or q) increases.

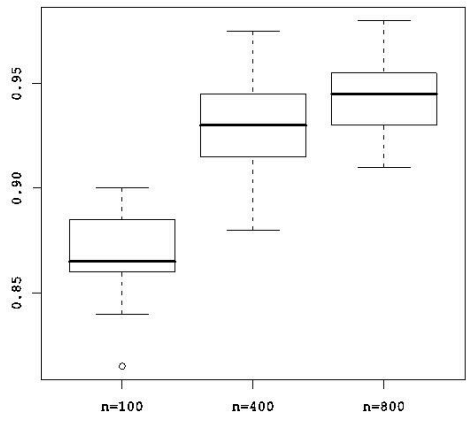

Figure 4.1 Box-plots for coverage probabilities of the nominal 95% prediction intervals of random effects in frailty models, according to increase of sample sizes.

In conclusion, for the inference of both fixed parameters and random effects in log-normal frailty models the h-likelihood method performs well as in Poisson HGLMs (Ha, 2008a). The prediction problems for random effects have been mainly studied in parametric HGLMs; see for example Lee and Nelder (1996), Ainsworth and Dean (2006) and Ha (2008a). However, the main focus of proposed method is to study a new h-likelihood prediction for random effects in semiparametric log-normal frailty models. Furthermore, it would be very interesting to extend to frailty models with various structures such as gamma distributed frailty (Hougaard, 2000), multi-component frailties (Ha et al., 2007) or non-PH structures (Ha, 2008b; Ha and MacKenzie, 2009).

References

Ainsworth, L. M. and Dean, C. B. (2006). Approximate inference for disease mapping. *Computational Statistics and Data Analysis,* **50**, 2552-2570.

Breslow, N. E. (1972). Discussion of Professor Cox's paper. *Journal of the Royal Statistical Society B,* **34**, 216-7.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society B,* **34**, 187-220.

Duchateau, L. and Janssen, P. (2008). *The frailty model,* New York: Springer-Verlag.

Ha, I. D. (2008a). On estimation of random effect in Poisson HGLMs. *Journal of Korean Data & Information Science Society*, **19**, 375-383.

Ha, I. D. (2008b). Frailty survival models with a non-PH function. *Journal of Korean Data & Information Science Society*, **19**, 343-351.

Ha, I. D. and Lee, Y. (2003). Estimating frailty models via Poisson hierarchical generalized linear models. *Journal of Computational Graphical Statistics,* **12**, 663-681.

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

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

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

Ha, I. D. and MacKenzie, G. (2009). Robust frailty modelling using non-proportional hazards models. *Statistical Modelling,* in press.

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

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

Lee, Y., Nelder, J. A. and Pawitan, Y. (2006). *Generalised linear models with random effects: unified analysis via h-likelihood,* London: Chapman and Hall.

McGilchrist, C. A. (1993). REML estimation for survival models with frailty. *Biometrics,* **49**, 221-225.

Ripatti, S. and Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics,* **56**, 1016-1022.

Vaida, F. and Xu, R. (2000). Proportional hazards model with random effects. *Statistics in Medicine*, **19**, 3309-3324.