Journal of Biometrics & Biostatistics
ISSN: 2155-6180
Sinha, J Biomet Biostat 2012, S1
DOI: 10.4172/2155-6180.S1-008
Research Article Open Access

The Use of Score Tests for Frailty Variance Components in Recurrent Event Data

Sanjoy K. Sinha*

School of Mathematics and Statistics, Carleton University, Ottawa, ON, Canada

Abstract

In the analysis of recurrent event data, frailties are commonly used to model the dependence structure among repeated event times within an individual. Often it is of interest to test whether the variance component in a frailty model is zero. It is well-known that the usual asymptotic mixtures of chi-square distributions of the score statistics for testing constrained variance components do not necessarily hold. In this paper, we propose and explore a stochastic permutation score test based on randomly permuting the indices associated with the individuals of a survival model. An empirical study suggests that the proposed score test has approximately the correct level of significance and is more powerful than the asymptotic score test based on the mixture of chi-square distributions. The proposed test is illustrated using two sets of actual recurrence failure time data obtained from clinical experiments.

Keywords: Failure time data; Frailty; Maximum likelihood; Permutation test; Proportional hazards model; Score test; Variance component

Introduction

In many applications of survival data, often the survival times are not independent. Such data can arise when different individuals share some common feature or when we observe recurrent event times within an individual. Frailties may be used to model the dependence structure in such survival data. A frailty can also be used to describe an unobservable genetic effect if individuals are in sibling groups or an environmental effect if individuals are grouped by households. Some experiments lead to repeated event times within an individual, and in such cases, frailties can be used to model the association among the repeated outcomes.

Frailties can also be used to model survival data in the presence of unobserved heterogeneity. For univariate (independent) failure times, these may be used to describe the effects of unobserved covariates in a proportional hazards model. For multivariate (dependent) failure times, these may be used to model the dependence structure among the multivariate observations, where it is usually assumed that given the frailty the multivariate failure times are conditionally independent. Frailty models are extensions of the Cox regression models [1], and provide an alternative way of modelling survival data where the hazard function is not monotonic, or where the hazards are not proportional. Frailties are also used to model survival times for grouped individuals, such as twins or family members, and recurrence survival times for the same individual. It is usually assumed that the frailties are random observations from a probability distribution with mean zero, but with an unknown variance component that needs to be estimated. A number of authors studied frailty models for describing heterogeneity in survival data, which include McGilchrist and Aisbett [2]; Klein [3]; McGilchrist [4]; Aalen [5,6]; Hougaard [7]; Klein and Moeschberger [8]; and Hougaard [9].

Often we are interested in testing whether the individuals in recurrent event data or groups in clustered survival data are homogeneous for given explanatory variables, or equivalently, whether the variance component in a frailty model is zero. In this paper, we investigate a score test for testing the homogeneity of the individuals (or groups). The score statistic is derived from a Taylor series expansion of the likelihood function about the mean of the frailty. The development of the test procedure is computationally less intensive as compared to the full likelihood ratio test in that it only requires the calculation of the score statistic under the null hypothesis of no frailty. It is well-known that for finite samples, the usual one-sided score tests based on mixtures of chi-squares often result in incorrect estimates of the level of significance (see, for example, Shephard and Harvey [10]; Shephard [11]; Pinheiro and Bates [12]; Crainiceanu, Ruppert and Vogelsang [13]; Crainiceanu and Ruppert [14]; Fitzmaurice, Lipsitz and Ibrahim [15]; and Sinha [16]). As a remedy, here we propose and explore a permutation test that approximates the *p*-value of the one-sided score test for the variance component. The proposed test provides approximately the correct level of significance under the null hypothesis even for small samples, and is also more powerful than tests based on mixtures of chi-square distributions.

The paper is organized as follows. Section "The Score Test" describes the proposed score statistic for testing the variance component in a frailty model. It also describes the permutation method for approximating the p-value of the score test. Section "Illustrative Example" presents an example using a survival model with a shared frailty for recurrent event data, and illustrates the calculation of the score statistic for testing the variance component of the frailty term. Section "Simulation Study" presents results from a simulation study that was carried out to investigate the finite-sample properties of the proposed permutation score test as well as the asymptotic score test based on the mixture of chi-square distributions. Section "Applications" provides applications of the proposed test with two sets of actual recurrence failure time data obtained from clinical experiments. Section "Discussion" concludes the paper with some discussion.

The Score Test

Suppose, we observe recurrent event data from N individuals,

***Corresponding author:** Sanjoy K. Sinha, School of Mathematics and Statistics, Carleton University, Ottawa, ON, K1S 5B6 Canada, E-mail: sinha@math.carleton.ca

Received July 07, 2012; **Accepted** August 29, 2012; **Published** September 03, 2012

Citation: Sinha SK (2012) The Use of Score Tests for Frailty Variance Components in Recurrent Event Data. J Biomet Biostat S1:008. doi:10.4172/2155-6180.S1-008

where the i^{th} individual has n_i repeated event times. Let T_{ij} denote the j^{th} event time for the i^{th} individual and δ_{ij} denote the censoring information (δ_{ij} = 0 if T_{ij} is right-censored; δ_{ij} = 1 otherwise). We assume that the censoring is non-informative. Let x_{ij}= (x_{ij1},..., x_{ijp})' denote the vector of explanatory variables associated with the (i,j)th event. Suppose, conditional on the frailty u_i, the hazard rate for the j^{th} event is of the form:

$$h_{ij}(t) = h_0(t)\exp(x'_{ij}\beta + u_i) \tag{1}$$

where $h_0(t) = h_0(t;\theta)$ is a baseline hazard function depending on a vector of unknown parameters θ and β is a vector of unknown regression coefficients. We assume that the frailties u_is are independent and follow a normal distribution with mean 0 and variance component σ_u^2. As the hazard rates within an individual share the same frailty, the event times within an individual are not independent. However, in the limit as $\sigma_u^2 \to 0$, they tend to be independent.

Given the data $\{(t_{ij},\delta_{ij});i=1,...,N;j=1,...,n_i\}$, the marginal likelihood of the model parameters may be expressed in the form:

$$\begin{aligned} L(\beta,\theta,\sigma_u^2) &= \prod_{i=1}^{N}\int_{u_i}\prod_{j=1}^{n_i}\{f_{ij}(t_{ij})\}^{\delta_{ij}}\{S_{ij}(t_{ij})\}^{1-\delta_{ij}} f(u_i)du_i \\ &= \prod_{i=1}^{N}\int_{u_i}\prod_{j=1}^{n_i}\{h_{ij}(t_{ij})\}^{\delta_{ij}} S_{ij}(t_{ij}) f(u_i)du_i \\ &= \prod_{i=1}^{N} E_{u_i}\left\{\prod_{j=1}^{n_i} g(t_{ij}\,|\,u_i)\right\}, \end{aligned} \tag{2}$$

where $g(t_{ij}\,|\,u_i) = \{h_{ij}(t_{ij})\}^{\delta_{ij}} S_{ij}(t_{ij})$ with $S_{ij}(t)$ being the survivor function for the (i,j)th event at time t, $f_{ij}(t)$, is also the corresponding density function at time t, and E_{u_i} denotes the expectation with respect to the distribution of the frailty u_i.

Similarly to Cox [17] (see also, Dean [18]), we expand the term $\prod_j g(t_{ij}\,|\,u_i) \equiv g^*(t_i\,|\,u_i)$ in (2) using a Taylor series expansion about $E_{u_i}(u_i) = 0$, and take expectations with respect to u_i to obtain the marginal likelihood for the i^{th} individual as

$$L_i(\beta,\theta,\sigma_u^2) = g^*(t_i\,|\,u_i = 0) + \sum_{k=2}^{\infty}\frac{a_k}{k!}\left\{\frac{\partial^k}{\partial u_i^k} g^*(t_i\,|\,u_i)\Big|_{u_i=0}\right\}, \tag{3}$$

where $a_k = E_{u_i}(u_i^k)$.

To test the homogeneity of individuals, we set the null hypothesis that there is no difference in event times between individuals for the given explanatory variables, whereas the alternative hypothesis is that the event times for a given individual share a common frailty. This is equivalent to testing the null hypothesis $H_0 : \sigma_u^2 = 0$, against the one-sided alternative hypothesis $H_0 : \sigma_u^2 > 0$. The score test for testing the null is based on the score function

$$U(\gamma) = \sum_{i=1}^{N}\frac{\partial \log L_i(\gamma,\sigma_u^2)}{\partial\sigma_u^2}\Big|_{\sigma_u^2=0} = \sum_{i=1}^{N} S_i^*(\gamma) \tag{4}$$

where $\gamma = (\beta',\theta')'$ and $S_i^*(\gamma)$ is the score function of σ_u^2 for the i^{th} individual evaluated as $\sigma_u^2 = 0$:

$$\begin{aligned} S_i^*(\gamma) &= \frac{\partial \log L_i(\gamma,\sigma_u^2)}{\partial\sigma_u^2}\Big|_{\sigma_u^2=0} \\ &= \frac{1}{2}\left\{\left(\sum_{j=1}^{n_i}\frac{\partial}{\partial u_i}\log g(t_{ij}\,|\,u_i)\right)^2 + \sum_{j=1}^{n_i}\frac{\partial^2}{\partial u_i^2}\log g(t_{ij}\,|\,u_i)\right\}\Big|_{u_i=0} \\ &= \frac{1}{2}\left\{\left(\sum_{j=1}^{n_i}\{\delta_{ij} - H_0(t_{ij})\exp(x'_{ij}\beta)\}\right)^2 + \sum_{j=1}^{n_i} H_0(t_{ij})\exp(x'_{ij}\beta)\right\} \end{aligned} \tag{5}$$

with $H_0(t) \equiv H_0(t;\theta)$ being the baseline cumulative hazard function at time t, depending on the parameters θ.

The score statistic is defined as a function of $U_0 \equiv U(\tilde{\gamma})$, where $\tilde{\gamma} = (\tilde{\beta}',\tilde{\theta}')'$ are the ML estimators of the nuisance parameters $\gamma = (\beta',\theta')'$ under $H_0 : \sigma_u^2 = 0$. An approximate variance of the score function $U(\gamma)$ can be derived from the observed Fisher information matrix:

$$I(\gamma) = \begin{pmatrix} I_{\gamma\gamma} & I_{\gamma\sigma} \\ I_{\sigma\gamma} & I_{\sigma\sigma} \end{pmatrix},$$

evaluated at $\sigma_u^2 = 0$, where

$$I_{\gamma\gamma} = -\sum_{i=1}^{N}\frac{\partial^2 \log L_i(\gamma,\sigma_u^2)}{\partial\gamma\partial\gamma'}\Big|_{\sigma_u^2=0}$$

$$I_{\gamma\sigma} = I'_{\sigma\gamma} = -\sum_{i=1}^{N}\frac{\partial^2 \log L_i(\gamma,\sigma_u^2)}{\partial\gamma\partial\sigma_u^2}\Big|_{\sigma_u^2=0}$$

and

$$I_{\sigma\sigma} = -\sum_{i=1}^{N}\frac{\partial^2 \log L_i(\gamma,\sigma_u^2)}{\partial(\sigma_u^2)^2}\Big|_{\sigma_u^2=0}$$

For calculating the Fisher information $I(\gamma)$, after some algebra, we can show that for the i^{th} individual,

$$\frac{\partial^2 \log L_i(\gamma,\sigma_u^2)}{\partial\beta\partial\beta'}\Big|_{\sigma_u^2=0} = -\sum_{j=1}^{n_i} H_0(t_{ij})\exp(x'_{ij}\beta)x_{ij}x'_{ij},$$

$$\frac{\partial^2 \log L_i(\gamma,\sigma_u^2)}{\partial\theta\partial\beta'}\Big|_{\sigma_u^2=0} = -\sum_{j=1}^{n_i}\frac{\partial H_0(t_{ij})}{\partial\theta}\exp(x'_{ij}\beta)x'_{ij},$$

$$\frac{\partial^2 \log L_i(\gamma,\sigma_u^2)}{\partial\theta\partial\theta'}\Big|_{\sigma_u^2=0} = \sum_{j=1}^{n_i}\left\{\delta_{ij}\frac{\partial^2 \log h_0(t_{ij})}{\partial\theta\partial\theta'} - \frac{\partial^2 H_0(t_{ij})}{\partial\theta\partial\theta'}\exp(x'_{ij}\beta)\right\},$$

$$\frac{\partial^2 \log L_i(\gamma,\sigma_u^2)}{\partial\beta\partial\sigma_u^2}\Big|_{\sigma_u^2=0} = -\left\{\sum_{j=1}^{n_i} H_0(t_{ij})\exp(x'_{ij}\beta)x_{ij}\right\}\left\{\sum_{j=1}^{n_i}\delta_{ij} - \sum_{j=1}^{n_i} H_0(t_{ij})\exp(x'_{ij}\beta) + \frac{1}{2}\right\},$$

$$\frac{\partial^2 \log L_i(\gamma,\sigma_u^2)}{\partial\theta\partial\sigma_u^2}\Big|_{\sigma_u^2=0} = -\left\{\sum_{j=1}^{n_i}\frac{\partial H_0(t_{ij})}{\partial\theta}\exp(x'_{ij}\beta)\right\}\left\{\sum_{j=1}^{n_i}\delta_{ij} - \sum_{j=1}^{n_i} H_0(t_{ij})\exp(x'_{ij}\beta) + \frac{1}{2}\right\}$$

and

$$\frac{\partial^2 \log L_i(\gamma,\sigma_u^2)}{\partial(\sigma_u^2)^2}\Big|_{\sigma_u^2=0} = -\left\{\sum_{j=1}^{n_i} H_{ij}(t_{ij})\right\}\left\{\sum_{j=1}^{n_i}\delta_{ij} - \sum_{j=1}^{n_i} H_{ij}(t_{ij})\right\}\left\{\sum_{j=1}^{n_i}\delta_{ij} - \sum_{j=1}^{n_i} H_{ij}(t_{ij}) + 1\right\} + \frac{1}{2}\left\{\sum_{j=1}^{n_i} H_{ij}(t_{ij})\right\}^2 - \frac{1}{4}\sum_{j=1}^{n_i} H_{ij}(t_{ij})$$

where $H_{ij}(t) = H_0(t)\exp(x'_{ij}\beta)$.

The variance of the score function $U(\gamma)$ may be approximated from

$$D(\gamma) = I_{\sigma\sigma} - I_{\sigma\gamma} I_{\gamma\gamma}^{-1} I_{\gamma\sigma} . \quad (6)$$

To test the null H_0 against the one-sided alternative H_1, following Silvapulle and Silvapulle [19], we use a score statistic in the form

$$T = \frac{U_0^2}{D_0} - \inf\left\{\frac{(U_0 - \delta)^2}{D_0} : \delta > 0\right\}, \quad (7)$$

where $D_0 = D(\tilde{\gamma})$. The development of this statistic is motivated by the fact that in the limit as $N\to\infty$, it becomes the likelihood ratio statistic (see Silvapulle and Silvapulle [19], for details). For positive $\hat{\sigma}_u^2$, the score at zero is positive, and the infimum in (7) becomes zero in {δ>0}. But when $\hat{\sigma}_u^2$ is negative, the score at zero is also negative, and so the infimum in (7) is attained at δ=0 and the statistic T becomes zero. As noted by Verbeke and Molenberghs [20], there should be valid models for sufficiently small but negative values of $\hat{\sigma}_u^2$, even under a constrained setting.

As we noted earlier, under the null $H_0 : \sigma_u^2 = 0$, the score statistic T for the one-sided test $H_0 : \sigma_u^2 > 0$ does not follow the usual chi-square distribution, since the value of $\hat{\sigma}_u^2$ under H_0 is on the boundary of the parameter space. As the number of individuals N tends to ∞, T has the mixture distribution: $0.5\chi_0^2 + 0.5\chi_1^2$, where χ_0^2 has a point mass at 0 and χ_1^2 has a chi-square distribution with one degree of freedom. However, for finite values of N, this mixture of chi-squares may lead to incorrect level of significance. In the case of a linear mixed model with a fixed intercept and a random group (or cluster) effect, Crainiceanu and Ruppert [14] showed that for fixed N and $n_i \to \infty$, the likelihood ratio statistic has a mixture distribution in the form $(1-\alpha_N)\chi_0^2 + \alpha_N\chi_1^2$, where α_N is determined by the group size N. However, for survival models with frailties, the appropriate mixture of chi-squares may not be straightforward to obtain.

To approximate the distribution of the score statistic, we propose a permutation method. The proposed permutation score test provides approximately the correct level of significance under the null, even for small samples. To obtain an approximate p-value of the score test based on the permutation method, we adopt the following algorithm:

1. For given survival data, obtain estimates $\tilde{\gamma}$ of the nuisance parameters γ under the null hypothesis $H_0 : \sigma_u^2 = 0$. Using equations (4)-(7), compute the observed value of the score statistic T, denoted by T_{obs}.
2. Hold the number of events n_i fixed for the ith individual, and then permute the individual indices of the given data at random. Compute the score statistic T based on the permutation sample. Produce a series of test statistics $(T^1,...,T^R)$ for R permutation samples by repeating this step a large number of times R.
3. The approximate p-value of the stochastic permutation score test is determined by the proportion of permutation samples with $T^r \geq T_{obs}$.

Note that Fitzmaurice, Lipsitz and Ibrahim [15] used a similar permutation method to approximate the p-value of a likelihood ratio test for testing the variance components in multilevel generalized linear mixed models. Here, we adopt the permutation method to approximate the p-value of the proposed score test for testing a frailty variance component in recurrent event data.

We have described the score test for a frailty variance component in the framework of recurrent event data. This approach can also be used for testing the significance of frailties in clustered survival data. We may encounter such clustered data when different individuals share some common characteristics. For example, in a multicenter clinical experiment, the survival times of patients from the same center may be more similar as compared to those from different centers. This could be due to different health care services provided to the patients in the different centers. Here, we can treat the centers as clusters and describe the homogeneity of survival times of the patients within a center using a shared frailty model.

Illustrative Example: A Proportional Hazards Model with a Shared Frailty

Consider a simple two-level exponential hazards model with two covariates x_{ij1} and x_{ij2}, and with a single frailty u_i:

$$h_{ij}(t) = h_0(t)\exp(\beta_1 x_{ij1} + \beta_2 x_{ij2} + u_i), i = 1,...,N; j = 1,...,n , \quad (8)$$

where $h_0(t)=\lambda$ is the baseline hazard function, and the frailties u_i are assumed to be independently and normally distributed with mean zero and unknown variance component σ_u^2. Model (8) may be rewritten in the form

$$h_{ij}(t) = \exp(\beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + u_i) = \exp(x_{ij}'\beta + u_i) ,$$

where $\beta_0 = \log\lambda$, $x_{ij} = (1, x_{ij1}, x_{ij2})'$ represents the vector of covariates, and $\beta = (\beta_0, \beta_1, \beta_2)'$ represents the vector of unknown regression coefficients.

For model (8), the score function (4) takes the form

$$U(\beta) = \frac{1}{2}\sum_{i=1}^{N}\left\{\left(\sum_{j=1}^{n}\{\delta_{ij} - t_{ij}\exp(x_{ij}'\beta)\}\right)^2 - \sum_{j=1}^{n} t_{ij}\exp(x_{ij}'\beta)\right\}. \quad (9)$$

Also, for the Fisher information matrix, we have under $H_0 : \sigma_u^2 = 0$:

$$I_{\beta\beta} = \sum_{i=1}^{N}\sum_{j=1}^{n} t_{ij}\exp(x_{ij}'\beta)x_{ij}x_{ij}'$$

$$I_{\beta\sigma} = I_{\sigma\beta}' = \sum_{i=1}^{N}\left\{\sum_{j=1}^{n} t_{ij}\exp(x_{ij}'\beta)x_{ij}\right\}\left\{\sum_{j=1}^{n}\delta_{ij} - \sum_{j=1}^{n} t_{ij}\exp(x_{ij}'\beta) + \frac{1}{2}\right\},$$

and

$$I_{\sigma\sigma} = \sum_{i=1}^{N}\left\{\sum_{j=1}^{n} t_{ij}\exp(x_{ij}'\beta)\right\}\left\{\sum_{j=1}^{n}\delta_{ij} - \sum_{j=1}^{n} t_{ij}\exp(x_{ij}'\beta)\right\}$$
$$\left\{\sum_{j=1}^{n}\delta_{ij} - \sum_{j=1}^{n} t_{ij}\exp(x_{ij}'\beta) + 1\right\} - \frac{1}{2}\sum_{i=1}^{N}\left\{\sum_{j=1}^{n} t_{ij}\exp(x_{ij}'\beta)\right\}^2$$
$$+\frac{1}{4}\sum_{i=1}^{N}\sum_{j=1}^{n} t_{ij}\exp(x_{ij}'\beta)$$

The score statistic T takes the form (7) with the score function $U_0 = U(\tilde{\beta})$ and the variance function $D_0 = D(\tilde{\beta})$, where $D(\beta) = I_{\sigma\sigma} - I_{\sigma\beta}I_{\beta\beta}^{-1}I_{\beta\sigma}$ and $\tilde{\beta}$ is the ML estimator of β under $H_0 : \sigma_u^2 = 0$. We performed a simulation study to investigate the empirical properties of the proposed permutation score test for the variance component in (8). Details are provided in the next section.

Simulation Study

To study the finite-sample properties of the proposed permutation test, we ran a series of simulations. The "true" event times y_{ij} with

frailties were generated from the hazards model (8) with the values of the regression coefficients being fixed at $\beta_0 = \log(\lambda) = \log(0.08)$, $\beta_1 = 0.5$ and $\beta_2 = -0.25$. For each combination of N=50, 100, 200 individuals and n=2, 10, 50 repeated event times within individuals, we performed a simulation study based on 1000 replicates of data sets. The censoring times c_{ij} were assumed to be independently and identically distributed exponential random variables with mean $1/\lambda^* = 1/0.08$, and the censoring mechanism was assumed to be non-informative. The observed data were $\{(t_{ij},\delta_{ij});i=1,...,N;j=1,...,n\}$, where $t_{ij}=\min(y_{ij},c_{ij})$ and $\delta_{ij}=I(y_{ij}< c_{ij})$.

The value of the frailty variance component σ_u^2 in (8) was set to zero when investigating the empirical level of significance of the proposed test. We compared the p-value of the permutation test to the asymptotic p-value obtained by the (0.5, 0.5) mixture of chi-square distributions. The permutation p-value was based on B=1000 permutation samples, and the empirical level of the test was obtained as the proportion of samples for which the estimated p-values were less than the nominal level α=0.05.

Table 1 presents the estimated levels of significance of the two score tests. We note from the results that the level of the permutation test is generally much closer to the nominal 0.05 level of significance, as compared to the level based on the mixture of chi-square distributions. For the latter case, the levels get closer to the nominal 0.05 level only when the number of events n and the number of individuals N increase. The permutation test roughly provides the correct level of significance in each of the simulation configurations considered. Also, the approximate 95% normal confidence intervals of the true levels, based on the estimated standard errors, suggest that the empirical levels of the permutation tests are not significantly different from the nominal 0.05 level, whereas most of the levels from the asymptotic score tests are significantly different.

We also investigated the powers of the two score tests using the same simulation configurations as above. The empirical powers were calculated under the alternative hypothesis $H_1 : \sigma_u^2 = 0.25$. We used 1000 simulation replications for each simulation configuration, and also used 1000 permutation samples to find the permutation p-value of the score test. Table 2 presents the empirical powers of the two tests. It is clear that the proposed permutation test is generally more powerful than the test based on the mixture of chi-square distributions for the simulation configurations considered. For N=200 and n=2, however, the score test based on the asymptotic mixture appears to be more powerful than the permutation test. But it should be noted that the empirical powers of the asymptotic score test may not be reliable here as the test provides incorrect level of significance for the given sample size.

Individuals		Number of events (n)		
N	Test	2	10	50
50	Permutation	0.046 (0.0066)	0.047 (0.0067)	0.052 (0.0070)
	Mixture	0.005 (0.0022)	0.009 (0.0030)	0.016 (0.0040)
100	Permutation	0.045 (0.0066)	0.040 (0.0062)	0.052 (0.0070)
	Mixture	0.011 (0.0033)	0.019 (0.0043)	0.027 (0.0051)
200	Permutation	0.048 (0.0068)	0.052 (0.0070)	0.050 (0.0069)
	Mixture	0.015 (0.0038)	0.026 (0.0050)	0.031 (0.0055)

Table 1: Empirical level of significance of the score test for a frailty model (standard error in parenthesis).

Individuals		Number of events (n)		
N	Test	2	10	50
50	Permutation	0.168 (0.0118)	0.988 (0.0034)	1.000 (0.0000)
	Mixture	0.126 (0.0105)	0.976 (0.0048)	1.000 (0.0000)
100	Permutation	0.273 (0.0141)	1.000 (0.0000)	1.000 (0.0000)
	Mixture	0.257 (0.0138)	1.000 (0.0000)	1.000 (0.0000)
200	Permutation	0.457 (0.0158)	1.000 (0.0000)	1.000 (0.0000)
	Mixture	0.572 (0.0156)	1.000 (0.0000)	1.000 (0.0000)

Table 2: Empirical power of the score test for a frailty model (standard error in parenthesis).

The results from the simulation study demonstrate that the proposed permutation test generally has the correct level of significance and is also more powerful than the asymptotic test based on the mixture of chi-square distributions.

Applications

Bladder cancer data

Wei, Lin and Weissfeld [21] presented and analyzed some tumor recurrence data obtained from a bladder cancer study conducted by the Veterans Administration Cooperative Urological Research Group (see Byar [22] for details). In this study, all patients entering the trial had superficial bladder tumors. After removing these tumors transurethrally, patients were randomly assigned to one of three treatments: placebo, thiotepa and pyridoxine. During the study, many patients had multiple recurrences of tumors, and new tumors were removed at each visit. Due to the sparseness of the data, only the first four recurrence times were reported. One of the analyses considered by Byar [22] and Wei, Lin and Weissfeld [21] was based on the tumor recurrence times from patients in the two groups placebo and thiotepa.

Here, we revisit the tumor recurrence data, where we consider modelling the recurrence time of a patient measured from the removal of old tumors at a given visit until the recurrence of new tumors. The recurrence time T_{ij} represents the number of months from a given visit until the next j^{th} tumor recurrence for the ith patient (i=1,…,85; j=1,…,4). As before, δ_{ij} represents the censoring information (δ_{ij}=0 if T_{ij} is right-censored; δ_{ij}=1 otherwise). The covariates considered in the study are: $TREAT_i$=1 if the i^{th} patient is in the thiotepa group and 0 otherwise; $NUMBER_i$=number of initial tumors for the i^{th} patient; and $SIZE_i$=size of the largest initial tumor for the i^{th} patient. We consider a proportional hazards model with a shared frailty in the form

$$h_{ij}(t) = \exp(\beta_0 + \beta_1 TREAT_i + \beta_2 NUMBER_i + \beta_3 SIZE_i + u_i), \qquad (10)$$

for i=1,…,85;j=1,…,4, where t is the time from the beginning of the j^{th} recurrence interval and u_i is the random effect (frailty) for the i^{th} patient, assumed to be independently and normally distributed with mean 0 and variance component σ_u^2.

The null hypothesis to be tested is that there is no difference in recurrence times between subjects for the given explanatory variables ($H_0 : \sigma_u^2 = 0$), against the alternative that the recurrence times for the same individual share the same frailty ($H_1 : \sigma_u^2 > 0$). To perform the proposed score test, we first fit the model under the null $H_0 : \sigma_u^2 = 0$ using the maximum likelihood method. The score statistic produced a value of 12.001. Here, the permutation test based on 10000 permutation samples produced a p-value of 0.0005. The usual (0.5, 0.5) mixture of chi-squares also produced a small p-value of 0.00027, as expected. Clearly, both methods indicate strong evidence against the null, that is, there is significant difference in recurrence times between subjects for

the given explanatory variables. As σ_u^2 is significant, we fit the above hazards model with the frailty. The maximum likelihood estimates of the model parameters, and their corresponding approximate standard errors are shown in Table 3. Here, the treatment thiotepa appears to decrease the risk of recurrence of tumors, whereas this risk is increased by a large number of initial tumors.

Note, that as we compute the score statistics under the null hypothesis of no frailties, the permutation score test does not require much computation time even with a large number of permutation samples. For the above example, we found the permutation *p*-value of the score test based on 10000 permutation samples in about 4 minutes and 50 seconds using the **R** package on a 64-bit Operating System with AMD Turion(tm) II P540 Dual-Core Processor 2.40 GHz and with 4.00 GB RAM.

Kidney data

McGilchrist and Aisbett [2] published some recurrence data, and studied the recurrence of infection in kidney patients who were using a portable dialysis machine. The infection in patients occurs at the point of insertion of the catheter. When the infection occurs, the catheter must be removed, the infection cleared up, and then the catheter reinserted. Recurrence times are measured from insertion until the next infection. When the catheter is removed for other reasons, there is right censoring of the data. Also, the final recurrence time may be censored as each patient is followed for a predetermined number of recurrence times. The covariates considered in the study are: AGE, and binary indicators for FEMALE as well as disease types GN, AN and PKD.

We revisit the kidney data, and consider a proportional hazards model with a shared frailty in the form

$$h_{ij}(t) = \exp(\beta_0 + \beta_1 AGE_{ij} + \beta_2 FEMALE_i + \beta_3 GN_i + \beta_4 AN_i + \beta_5 PKD_i + u_i), \quad (11)$$

for i=1,…,38;j=1,2, where t is the time from the beginning of the j^{th} recurrence interval and u_i is the i^{th} patient effect (frailty), assumed to be independently and normally distributed with mean 0 and variance component σ_u^2.

Here, we are interested in testing if there is any significant difference in recurrence times between subjects for the given explanatory variables or, equivalently, if the recurrence times for the same individual share the same frailty. Initially, we conducted the score test based on a subset of the data with a fewer number of patients by considering the last 25 individuals (patients 14–38 in Table 1 of McGilchrist and Aisbett [2]) in order to investigate the performance of the two score tests under a small sample. The value of the score statistic is obtained as 0.0594. The asymptotic *p*-value based on the mixture of chi-squares provides a value of 0.4038, whereas the permutation test based on 10000 permutation samples produced a smaller *p*-value value of 0.0469. Clearly, unlike the asymptotic score test, the permutation method indicates that there is significant difference at 5% level in recurrence times between subjects for the given explanatory variables. This suggests that for small samples, the two test procedures can provide different conclusions, and the proposed permutation test may be preferable to the asymptotic score test in such a case.

In the next step, we performed the score test based on the recurrence times for all 38 patients. The score statistic produced a small value of 0.0339. The (0.5, 0.5) mixture of chi-squares produced a *p*-value of 0.4270, whereas the permutation test based on 10000 permutation samples produced a *p*-value value of 0.4115. Both methods indicates no evidence against the null that the subject-specific frailty is 0. So, we consider fitting the above hazards model with no frailty. The maximum likelihood estimates of the model parameters, and their corresponding asymptotic standard errors are presented in Table 4. Here the FEMALE group and the disease type PKD appear to have higher risk of recurrence of infection in kidney patients.

Coefficient	Estimate	SE	*z* value
INTERCEPT	–3.0478	0.5624	–5.42
TREAT	–0.6265	0.3239	–1.93
NUMBER	0.2485	0.0867	2.87
SIZE	–0.0176	0.1086	–0.16
σ_u^2	0.9221	0.3441	2.68

Table 3: Analysis of bladder cancer data using a proportional hazards model with a shared frailty.

Coefficient	Estimate	SE	*z* value
INTERCEPT	3.7055	0.4965	7.46
AGE	–0.0019	0.0112	–0.17
FEMALE	1.6118	0.3329	4.84
GN	–0.0580	0.4048	–0.14
AN	–0.5181	0.3911	–1.33
PKD	1.3264	0.5698	2.33

Table 4: Analysis of kidney data using a proportional hazards model with a shared frailty.

Discussion

For testing the significance of a variance component in a frailty model, the proposed permutation score test provides a simple alternative to computing the asymptotic *p*-values of score tests based on the (0.5, 0.5) mixture of chi-square distributions. Our limited simulation study suggests that the permutation test has approximately the correct level of significance, and is also more powerful than tests based on the mixture of chi-square distributions under finite samples. The permutation score test is easy to implement and is also attractive in that it only requires estimation of the the fixed effects parameters under the null hypothesis of no shared frailty in the proportional hazards model.

We have discussed the permutation test for testing homogeneity of individuals in two-level survival data. This test can be easily extended to multilevel survival data with more than two levels. For testing homogeneity of groups at a given level, we can permute the indices corresponding to that level. For example, consider the bladder cancer study discussed in Section "Applications". If the patients are nested within medical practices, then we can test for homogeneity at the practice level by permuting the practices that the patients are assigned to, while the recurrence times would remain with the same patients for a given practice.

Note that Sinha [16] studied a parametric bootstrap score test, based on random samples generated from the fitted model under the null, which approximates the *p*-value of a one-sided score test for variance components in generalized linear mixed models. This

parametric bootstrap procedure, however, may not be directly applicable to survival data, since the development of the test procedure is complicated by the fact that the survival times are often censored.

Acknowledgements

This research was partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

References

1. Cox DR (1972) Regression models and life tables. J R Stat Soc Series B (Methodological) 34: 187–220.
2. McGilchrist CA, Aisbett CW (1991) Regression with frailty in survival analysis. Biometrics 47: 461–466.
3. Klein JP (1992) Semiparametric estimation of random effects using the Cox model based on the EM algorithm. Biometrics 48: 795–806.
4. McGilchrist CA (1993) REML estimation for survival models with frailty. Biometrics 49: 221–225.
5. Aalen OO (1994) Effects of frailty in survival analysis. Stat Methods Med Res 3: 227–243.
6. Aalen OO (1998) Frailty models. In Statistical Analysis of Medical data: New Developments (Eds. Everitt BS and Dunn G), Arnold, London.
7. Hougaard P (1995) Frailty models for survival data. Lifetime Data Anal 1: 255–273.
8. Klein JP, Moeschberger ML (2003) Survival Analysis: Techniques for Censored and Truncated Data. Springer, New York.
9. Hougaard P (2000) Analysis of Multivariate Survival data. Springer, New York.
10. Shephard NG, Harvey AC (1990) On the probability of estimating a deterministic component in the local level model. J Time Ser Anal 11: 339–347.
11. Shephard N (1993) Maximum likelihood estimation of regression models with stochastic trend components. J Am Stat Assoc 88: 590–595.
12. Pinheiro JC, Bates DM (2000) Mixed-Effects Models in S and S-Plus. Springer-Verlag, New York.
13. Crainiceanu CM, Ruppert D, Vogelsang TJ (2002) Probability that the MLE of a variance component is zero with applications to likelihood ratio tests 1-21.
14. Crainiceanu CM, Ruppert D (2004) Likelihood ratio tests in linear mixed models with one variance component. J R Stat Soc Series B (Stat Methodol) 66: 165–185.
15. Fitzmaurice GM, Lipsitz SR, Ibrahim JG (2007) A note on permutation tests for variance components in multilevel generalized linear mixed models. Biometrics 63: 942–946.
16. Sinha SK (2009) Bootstrap tests for variance components in generalized linear mixed models. Can J Stat 37: 219–234.
17. Cox DR (1983) Some remarks on overdispersion. Biometrika 70: 269–274.
18. Dean CB (1992) Testing for overdispersion in Poisson and binomial regression models. J Am Stat Assoc 87: 451–-457.
19. Silvapulle MJ, Silvapulle P (1995) A score test against one-sided alternatives. J Am Stat Assoc 90: 342–349.
20. Verbeke G, Molenberghs G (2003) The use of score tests for inference on variance components. Biometrics 59: 254–262.
21. Wei LJ, Lin DY, Weissfeld L (1989) Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. J Am Stat Assoc 84: 1065–1073.
22. Byar DP (1980) The veterans administration study of chemoprophylaxis for recurrent stage I bladder tumors: comparisons of placebo, pyridoxine, and topical thiotepa. In Bladder Tumors and Other Topics in Urological Oncology, eds. M. Pavone-Macaluso, P. H. Smith, and F. Edsmyn. Plenum, New York 363–370.

This article was originally published in a special issue, **Advances in Markov Chain Monte Carlo Methods and Survival Analysis** handled by Editor(s). Dr. Faming Liang, Texas A&M University, USA; Dr. Nengjun Yi, University of Alabama at Birmingham, USA; Dr. Wenqing He, University of Western Ontario, Canada; Dr. Liuquan Sun, Institute of Applied Mathematics, Academy of Mathematics and Systems Science, China