A SAS MACRO FOR THE POSITIVE STABLE FRAILTY MODEL

Youyi Shu and John P. Klein, Medical College of Wisconsin
Youyi Shu, Biostatistics, 8701 Watertown Plank Road, Milwaukee, WI 53226

Key Words: survival analysis, frailty model, EM algorithm, SAS macro.

Abstract:

A SAS macro to extend the Cox proportional hazards regression model to allow for positive stable frailties is presented. This macro computes, using a modified EM algorithm, estimates of the model parameters and their respective standard errors. The likelihood ratio test of the independence assumption is also provided. An example data set is used to illustrate the macro.

1 Introduction

Frailty or random effects models are useful in survival analysis for modeling associations between individuals in certain groups. For example, if one is interested in studying risk factors for a particular disease outcome or the effectiveness of some treatment, it is reasonable to believe that siblings who share a common genetic code and early environmental exposure will have event times more closely related than non-siblings. In human (or animal) studies, the family (or litter) forms natural groupings, and thus dependencies, between study subjects.

One way to model the dependence of the event times is through the introduction of a common random effect (either environmental or genetic), called a *frailty*. In this model, individuals in a natural group share an unobservable random covariate, W, which acts multiplicatively on the hazard rate of each group member. If the realization of W is greater than one, then all members of the group tend to experience the event of interest at an early time, while the opposite occurs if W is less than one. Hence a positive association between group members is induced by the frailty.

Although frailty models are becoming increasingly appealing in survival analysis, they are not widely used. One reason for this is the lack of any user-friendly software. This paper presents a SAS macro for positive stable frailty model, which is now available on our Web site: *http://www.biostat.mcw.edu/SoftMenu.html.*

Section 1 outlines the positive stable frailty model. Section 2 describes the SAS macro and its output. Section 3 gives an example to illustrate the macro, using data from Mantel et al. (1977) on a litter-matched tumorigenesis experiment. Finally, the discussion summarizes the positive stable frailty model and the SAS macro.

2 The Positive Stable Frailty Model

2.1 The Positive Stable Frailty Distribution

Suppose that we have data on the event times and covariate values of n individuals from some population. Our sample consists of $M_i \geq 1$ individuals from the ith subgroup of the population, $i = 1, \ldots, B$. Individuals within the ith subgroup have dependent event times due to some unobserved covariate information summarized in a frailty, W_i. Note that, in this formulation, subgroups of size one are allowed and in such a case that individual is affected by his or her own frailty.

For the jth individual in the ith subgroup, let X_{ij} denote the time to the event of interest. Let $\mathbf{Z}_{ij}$ be a vector of potential covariates associated with this individual. Suppose that, conditional on the frailty W_i, the hazard rate for this individual is of the form

$$\lambda(x|\mathbf{Z}_{ij}, W_i) = W_i \lambda_0(x) \exp(\boldsymbol{\beta}'\mathbf{Z}_{ij}), \qquad (1)$$

where $\lambda_0(x)$ is an arbitrary baseline hazard rate and $\boldsymbol{\beta}$ is a p-vector of unknown parameters. Notice that if $W_i = 1$ for all i, then the frailty model reduces to the usual Cox (1972) model for independent data.

Here, as suggested by Hougaard (1986), we assume that the W_i's are independent and identically distributed positive stable variates with Laplace transform

$$Lap(s) = E(e^{-sW}) = \exp(-s^{\theta}), \quad 0 < \theta \leq 1.$$

Small values of θ reflect greater heterogeneity between subgroups and thus a stronger association among subgroup members. The strength of association between two individuals, measured by Kendall's τ, is $(1-\theta)$, with $\theta = 1$ corresponding to independence between group members. Note that when $\theta = 1$ the W_i's are equal to 1 with probability one.

A traditional method of assessing the effects of risk factors is to examine the relative risk of experiencing the event of interest for an individual with covariate vector $\mathbf{Z}_1$ as compared to an individual with covariate vector $\mathbf{Z}_2$. From model (1), for two individuals within a subgroup (i.e., sharing a common value of frailty), the conditional relative risk is

$$RR_{\_within} = \exp\{\boldsymbol{\beta}'(\mathbf{Z}_1 - \mathbf{Z}_2)\}.$$

But in general, for two randomly selected members of the population with covariates $\mathbf{Z}_1$ and $\mathbf{Z}_2$, the unconditional relative risk is (Shu, 1997)

$$RR_{\_between} = \exp\{\theta\boldsymbol{\beta}'(\mathbf{Z}_1 - \mathbf{Z}_2)\}.$$

2.2 Semi-parametric Estimation Via the EM Algorithm

To estimate θ, $\boldsymbol{\beta}$, and cumulative baseline hazard rate $\Lambda_0(x) = \int_0^x \lambda_0(u)du$, a semi-parametric EM algorithm (Dempster et al., 1977) based on a profile likelihood for the conditional proportional hazards model is implemented in the SAS macro. Suppose our data, based on a sample of size n, consists of the triple $(T_{ij}, I_{ij}, \mathbf{Z}_{ij})$, $i = 1, \ldots, B$, $j = 1, \ldots, M_i$, where for the jth individual in the ith subgroup, T_{ij} is the time on study, I_{ij} is the event indicator ($I_{ij} = 1$ if the event has occurred; $I_{ij} = 0$ if the lifetime is right censored) and $\mathbf{Z}_{ij}$ is the vector of covariates or risk factors. To apply the EM algorithm, we treat the observed data as the incomplete data and the unobserved frailties, W_i's, as the missing information. First, note that for fixed θ, if we could observe the W_i's, the augmented log likelihood is equal to, up to a term free of the unknown parameters,

$$\begin{aligned} &L_1(\boldsymbol{\beta}, \Lambda_0; data, w_1, \ldots, w_B) \\ &= \sum_{i=1}^{B}\sum_{j=1}^{M_i}\Big\{ I_{ij}\left[\boldsymbol{\beta}'\mathbf{Z}_{ij} + \ln\lambda_0(T_{ij})\right] \\ &\quad - w_i\Lambda_0(T_{ij})\exp(\boldsymbol{\beta}'\mathbf{Z}_{ij})\Big\}. \qquad (2) \end{aligned}$$

The estimating algorithm proceeds by first making an initial guess at the values of $\boldsymbol{\beta}$ (and thus at $\Lambda_0(\cdot)$). The initial estimates are obtained by a standard Cox's program. To apply the E-step of the algorithm we need to compute the expected value of W_i, conditional on the observed data. Wang et al. (1995) showed that

$$E[W_i|data] = \frac{E\left[W_i^{D_i+1}\exp(-H_iW_i)\right]}{E\left[W_i^{D_i}\exp(-H_iW_i)\right]}, \quad i = 1, \ldots, B, \qquad (3)$$

where

$$H_i = \sum_{j=1}^{M_i}\Lambda_0(T_{ij})\exp(\boldsymbol{\beta}'\mathbf{Z}_{ij}), \qquad (4)$$

and $D_i = \sum_{j=1}^{M_i} I_{ij}$ is the observed number of deaths in the ith group. To evaluate (3) we have the following lemma from Wang et al. (1995):

Lemma 1 *If W follows a positive stable distribution, then*

$$E[W^q\exp(-sW)] = (\theta s^{\theta-1})^q\exp(-s^\theta)J[q,s], \quad q = 0, 1, \ldots; \; s > 0 \qquad (5)$$

where $J[q,s] = \sum_{m=0}^{q-1}\Omega_{q,m}s^{-m\theta}$ and $\Omega_{q,m}$ is a polynomial of degree m given recursively by

$$\begin{aligned} &\Omega_{q,0} = 1; \\ &\Omega_{q,m} = \Omega_{q-1,m} + \Omega_{q-1,m-1}\{(q-1)/\theta - (q-m)\}, \\ &\qquad\qquad m = 1, \ldots, q-2; \\ &\Omega_{q,q-1} = \theta^{1-q}\Gamma(q-\theta)/\Gamma(1-\theta). \end{aligned}$$

Using (3) and (5), we substitute $\hat{W}_i = E[W_i|data]$ into (2) for W_i in the E-step of the EM algorithm to obtain

$$\begin{aligned} &E_W[L_1(\boldsymbol{\beta}, \Lambda_0; data, w_1, \ldots, w_B)] \\ &= \sum_{i=1}^{B}\sum_{j=1}^{M_i}\Big\{ I_{ij}\left[\boldsymbol{\beta}'\mathbf{Z}_{ij} + \ln\lambda_0(T_{ij})\right] \\ &\quad - \hat{W}_i\Lambda_0(T_{ij})\exp(\boldsymbol{\beta}'\mathbf{Z}_{ij})\Big\}. \qquad (6) \end{aligned}$$

The M-step of the EM algorithm requires the maximization of (6) with respect to the unknown parameters, $\boldsymbol{\beta}$. To obtain the updated estimate of $\boldsymbol{\beta}$, note that (6) contains the nuisance cumulative baseline hazard rate, Λ_0. Using the profile likelihood construction technique proposed by Johansen (1983), if we first fix $\boldsymbol{\beta}$, Wang et al. (1995) showed that the semi-parametric estimator of $\Lambda_0(t)$ is

$$\hat{\Lambda}_0(t) = \sum_{T_{(k)} \le t}\frac{d_{(k)}}{\sum_{l\in R(T_{(k)})}\hat{W}_l\exp(\boldsymbol{\beta}'\mathbf{Z}_l)}, \qquad (7)$$

where $T_{(k)}$ is the kth smallest death time, regardless of subgroup (In the sequel, we shall denote $T_{(1)} < T_{(2)} < \cdots < T_{(D)}$ as the D distinct event times); $d_{(k)}$ is the number of deaths at $T_{(k)}$; $R(T_{(k)})$ is the set of individuals at risk at time $T_{(k)}$; $\hat{W}_l$ is the expected value of the frailty, given the data (See (3)); and $\mathbf{Z}_l$ is the covariate vector associated with the lth individual in the sample. Substituting (7) into (6) yields the profile log likelihood for $\boldsymbol{\beta}$, up to a term free of the parameter values,

$$\begin{aligned} & L_2(\boldsymbol{\beta}) \\ &= \sum_{k=1}^{D} \left\{ S_{(k)}\boldsymbol{\beta} - d_{(k)} \ln \left[\sum_{l \in R(T_{(k)})} \hat{W}_l \exp(\boldsymbol{\beta}'\mathbf{Z}_l) \right] \right\}, \end{aligned} \tag{8}$$

where $S_{(k)}$ is the sum of the covariate vectors of individuals who died at time $T_{(k)}$.

Iterating between the E and M steps until convergence yields an estimate of $\boldsymbol{\beta}(\theta)$ for this fixed value of θ. For these parameter values we then compute the full unaugmented likelihood given by (Wang et al., 1995)

$$\begin{aligned} & L_{Full}(\theta, \boldsymbol{\beta}(\theta), \Lambda_0; data) \\ &= \sum_{i=1}^{B} \Bigg\{ D_i \left[\ln\theta + (\theta - 1)\ln H_i \right] \\ &\quad - [H_i]^{\theta} + \ln\{J[D_i, H_i]\} \\ &\quad + \sum_{j=1}^{M_i} I_{ij} \left[\boldsymbol{\beta}(\theta)'\mathbf{Z}_{ij} + \ln\lambda_0(T_{ij}) \right] \Bigg\}. \end{aligned} \tag{9}$$

In the SAS macro, (9), which is a function of θ only, is maximized by a golden search technique (Press et al., 1992) after initially bracketing a maximum through a grid search method.

In summary, the estimation routine proceeds as follows:

Step 0. Using a modified Cox regression program, obtain initial estimates of $\boldsymbol{\beta}$ and Λ_0 from (8) and (7) respectively, with $\hat{W}_l = 1$ (i.e., $\theta = 1$).

Step 1. Fix θ. Using the current values of θ, $\boldsymbol{\beta}$, and Λ_0, compute $\hat{W}_l = E[W_l|data]$ from (3) and (5).

Step 2. Update the estimate of $\boldsymbol{\beta}$ (and Λ_0) using (8) (and (7)).

Step 3. Iterate between Steps 1 and 2 until convergence of $\boldsymbol{\beta}$.

Step 4. Repeat Steps 1-3 to construct the profile likelihood for θ using (9) and search for that value of θ which maximizes (9) using the search method mentioned above.

2.3 Standard Errors of the Estimates

In this section, we use the same approach proposed by Andersen et al. (1997) to estimate the standard errors of the model parameter estimates. To simplify notations, we denote the baseline hazard rate estimates as $\boldsymbol{\alpha} = (\alpha_{(1)}, \ldots, \alpha_{(D)})'$, where

$$\alpha_{(k)} = \hat{\lambda}_0(T_{(k)}) = \frac{d_{(k)}}{\sum_{l \in R(T_{(k)})} \hat{W}_l \exp(\boldsymbol{\beta}'\mathbf{Z}_l)}, \quad k = 1, \ldots, D. \tag{10}$$

Substituting (7) into (4), we get the estimator of H_i as:

$$\begin{aligned} H_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \sum_{j=1}^{M_i} \hat{\Lambda}_0(T_{ij}) \exp(\boldsymbol{\beta}'\mathbf{Z}_{ij}) \\ &= \sum_{j=1}^{M_i} \sum_{k=1}^{D} Y_{ij}(T_{(k)}) \alpha_{(k)} \exp(\boldsymbol{\beta}'\mathbf{Z}_{ij}), \end{aligned} \tag{11}$$

where $Y_{ij}(t) = I(T_{ij} \geq t)$.

Now substituting (10) and (11) into (9), we can rewrite the observed log likelihood as:

$$\begin{aligned} & L(\theta, \boldsymbol{\beta}, \boldsymbol{\alpha}) \\ &= \sum_{i=1}^{B} \Bigg[D_i \left\{ \ln\theta + (\theta - 1)\ln[H_i(\boldsymbol{\alpha}, \boldsymbol{\beta})] \right\} \\ &\quad - [H_i(\boldsymbol{\alpha}, \boldsymbol{\beta})]^{\theta} + \ln\left\{ J[D_i, H_i(\boldsymbol{\alpha}, \boldsymbol{\beta})] \right\} \Bigg] \\ &\quad + \sum_{i=1}^{B} \sum_{j=1}^{M_i} I_{ij} \boldsymbol{\beta}'\mathbf{Z}_{ij} + \sum_{k=1}^{D} d_{(k)} \ln[\alpha_{(k)}]. \end{aligned} \tag{12}$$

Note that L is a function of $(1 + p + D)$ parameters $(\theta, \beta_1, \ldots, \beta_p, \alpha_{(1)}, \ldots, \alpha_{(D)})$. Careful differentiation of (12) with respect to these parameters yields the observed information matrix. The estimated covariance matrix of θ and $(\beta_1, \ldots, \beta_p)$ is the upper left hand $(1+p) \times (1+p)$ submatrix of the full information matrix. Details can be found in Shu (1997).

3 The SAS Macro for the Positive Stable Frailty Model

The main part of the macro is the statistical analysis using the positive stable frailty model based on the EM algorithm. The final results are summarized in an analysis of variance table that will list the individual effects (including the dependence parameter and covariates) and each effect's degrees of freedom, maximum likelihood estimate, standard error, Wald test p-value, and relative risk (both within group and between group). Also reported are Kendall's τ and the last computed value of the log full likelihood. The likelihood ratio test of the independence assumption that $\theta = 1$ is performed automatically and the degrees of freedom, chi-square statistic and the corresponding p-value are printed.

Note that, for comparison's sake, the analysis of variance table and the log partial/full likelihood from the usual Cox independence model are printed first in the summary report of the final results.

To make the best use of the macro, there are several printing and output data set options that the user may wish to specify. The printing options include the report of grouping information, the iteration history and/or a summary table from it, the estimated variance-covariance matrix of $\hat{\theta}$ and $\hat{\boldsymbol{\beta}}$, and confidence limits for the relative risks (it's also possible to change the confidence coefficient for the relative risks). The output data set options are: (1) a data set containing $\hat{\theta}$, $\hat{\boldsymbol{\beta}}$, the estimated variance-covariance matrix of $\hat{\theta}$ and $\hat{\boldsymbol{\beta}}$, and the last computed value of the log full likelihood; (2) a data set containing ordered event times, baseline hazard rates, standard errors of the baseline hazard rates, and $\hat{\theta}$; and (3) a data set containing the grouping variable, survival time, censoring status, covariates, estimated linear predictor, cumulative baseline hazard rate, and $\hat{\theta}$.

4 Example

This section illustrates the macro through a simple example. The data used is from Mantel et al. (1977) who published a litter-matched tumorigenesis experiment with one drug treated rat and two placebo treated rats per litter, 50 female litters and 50 male litters. It is conceivable that the risk of tumor formation may depend on the genetic background shared within litters, but differing between litters. Thus there could be an intra-litter correlation in time to tumor appearance which is the event of interest. Death before tumor appearance implies censoring at the time of death.

We perform a semi-parametric analysis assuming positive stable frailties for all the 100 litters (including both female and male). The two covariates are DRUG and SEX, which are the indicators of treatment group and male rat, respectively. It will be seen from the results that both the treatment and sex are significant (with p-values 0.0108 and 0.0001, respectively) while the dependence parameter is not significant (p-value 0.5663).

The following SAS code demonstrates the simplest possible macro invocation with no options.

```
/* Creating the data set */
data mydata;
infile 'mantel.dat';
input litter time censor drug sex;
run;

/* Macro loading */
%include 'ps_frail.macro';

/* Macro invocation */
%ps_frail(mydata)
```

The main SAS output from the example:

```
   ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
                   Summary  of  Final  Results
   ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
          /*******************************************/
          /*       Usual Cox Independence Model      */
          /*******************************************/
Log partial likelihood = -200.4718
Log full likelihood = -228.9487

             Analysis of Maximum Likelihood Estimates

     VARIABLE  DF   Estimate     Stderr     p_Wald         RR
     DRUG       1     0.7848     0.3093     0.0112     2.1920
     SEX        1    -3.0626     0.7248     0.0001     0.0468
```

```
          /*********************************************/
          /*      Positive Stable Frailty Model        */
          /*********************************************/
Dependence Parameter: THETA= 0.9497
Kendall's TAU = 0.0503
Log full likelihood = -228.7531

    Likelihood Ratio Test of Independence Model (H0: THETA=1)
                 DF     Chi-square       p-value
                  1         0.3912        0.5317

                       Analysis of Maximum Likelihood Estimates

EFFECT        DF   Estimate     Stderr     p_Wald  RR_within  RR_between
Depend Parm    1     0.9497     0.0876     0.5663          .           .
DRUG           1     0.8023     0.3146     0.0108     2.2306      2.1425
SEX            1    -3.1808     0.7973     0.0001     0.0416      0.0488
```

5 Discussion

The positive stable frailty model is gaining popularity in survival analysis. This SAS macro has made the model more accessible and certainly many people will take advantage of it. The macro calculates the model parameter estimates and respective standard errors and the confidence limits for the relative risks. It also performs a likelihood ratio test on the dependence parameter to see whether we need the frailty model or just Cox's independence model for a particular study. Nevertheless, the statistical analyses for both models are reported. It is worth mentioning that the results from Cox independence model are the same as those of PROC PHREG (SAS/STAT *Changes and Enhancements through Release 6.11*), and that the model parameter estimates from the positive stable frailty model are essentially identical with those presented in Wang et al. (1995).

It should also be pointed out that the variance calculations in Wang et al. (1995) are incorrect, since the variability of the estimates of the baseline hazard rates was not taken into account. This macro uses a technique proposed by Andersen et al. (1997) to obtain the variance estimates of the model parameters.

This macro is mainly written in SAS/IML, which is a high level programming language and is relatively slow. For some large applications, extensive computer time may be required. For the example in Section 4, the CPU time needed is 2 minutes 8 seconds on our Solaris 2.0 system. If we use only the data for 50 female litters (leaving out the covariate SEX in the model then), the CPU time will reduce to 1 minute 35 seconds.

Acknowledgement

This research was supported by Grant R01-CA54706-07 from the National Cancer Institute.

References

[1] Andersen, P.K., Klein, J.P., Knudsen, K.M., and Palacios, R.T.Y. (1997). Estimation of variance in Cox's regression model with shared Gamma frailties. *Biometrics* **53**, 1475-1484.

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

[3] Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39**, 1-38.

[4] Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions. *Biometrika* **73**, 387-396.

[5] Johansen, S. (1983). An extension of Cox's regression model. *International Statistical Review* **51**, 258-262.

[6] Mantel, N., Bohidar, N.R., and Ciminera, J.L. (1977). Mantel-Haenszel analyses of litter-matched time-to-response data, with modifications for recovery of interlitter information. *Cancer Research* **37**, 3863-3868.

[7] Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. (1992). *Numerical Recipes in C, Second Edition.* Cambridge University Press, New York.

[8] Shu, Y. (1997). *A SAS Macro for the Positive Stable Frailty Model.* Master's Thesis, Medical College of Wisconsin, Milwaukee, Wisconsin.

[9] Wang, S.T., Klein, J.P., and Moeschberger, M.L. (1995). Semi-parametric estimation of covariate effects using the positive stable frailty model. *Applied Stochastic Models and Data Analysis* **11**, 121-133.