9. Introduction to Multivariate Survival Data Analysis

Multivariate survival data are commonly encountered in many application areas, in which study subjects may experience multiple events or failures, or the event times of subjects may be grouped or clustered. Conventionally, people refer to the former situation as multiple events data and the latter as clustered events data. A very special case of multiple events data is recurrent events data, in which subjects often experience repeated occurrences of the same type of event.

Examples of multivariate survival data

1.a. The famous bladder tumor recurrence data (Byar, 1980). The data set was obtained from a randomized clinical trial assessing the effect of treatment thiotepa on the recurrence of bladder tumors. There were 38 patients in the thiotepa group with a total of 45 observed recurrence times, and 48 placebo patients with a total 87 observed recurrences. Beside the treatment assignment, two other baseline covariates: number of initial tumors and size of the largest initial tumor, are also recorded. The goal is to study how the treatment assignment and the two baseline covariates affect the timing and frequency of the bladder tumor recurrence.

1.b. Infection occurrence times among leukemia patients receiving bone marrow transplants (Pepe and Cai, 1993); Times to inpatient hospital admission among intravenous drug users (Wang et al., 2001); Repeated attacks of asthma and epileptic seizures; among many other medical studies.

1.c. Times to warranty claims for a particular car model (Kalbfleisch et al., 1991); Times to return to prisons of convicts after the first release; Times of buying a certain brand of product from potential customers; among many other applications in economics, sociology and software engineering.

1.d. Women's health initiative clinical trial (Women's health initiative study group, 1998). In a study of dietary component, over 48,000 postmenopausal American women were randomized to either a low-fat eating pattern or to control status. The primary endpoints

of interest are occurrence times of breast cancer and colorectal cancer and the secondary outcomes are the occurrence times of coronary heart disease. Therefore, a study subject may have different types of event times.

2.a. The the diabetic retinopathy study (Huster et al., 1989). The dataset contains 197 patients, who were a 50% random sample of the patients with "high-risk" diabetic retinopathy as defined by the Diabetic Retinopathy Study (DRS). Each patient had one eye randomized to laser treatment and the other eye served as an untreated control. For each eye, the event of interest was the time from initiation of treatment to the severe visual loss (call it "blindness"). Besides treatment indicator (1 = treatment; 0 = control), four other covariates: laser type (0 = xenon, 1 = argon), age at diagnosis of diabetics, type of diabetics (0= juvenile, 1 = adult) and risk group (6-12), are also included. In this example, the two event times of the same subject may be highly correlated.

2.b. The familial aggregation of breast cancer. It is well known that both genetic and environmental factors are important for breast cancer. To design a study, a number of families are collected. Baseline information on probands, relatives (parents, aunts and uncles, sisters and brothers, sons and daughters) are obtained by interviews, followup letters and telephone calls. Familial members will be followed for occurrence of breast cancer and for marry-ins to the families, the followup starts at the age that they married into the family tree. Other examples include the famous Australian twin study (Duffy et al., 1990).

Data notations

- for recurrent events data

  1. time to events: let T_{ik}, $i = 1, \cdots, n$, $k = 1, \cdots, m_i$, be the time to kth recurrence of subject i, where m_i is the total number of events taken on subject i. Let C_i be the censoring time and Z_i the covariate information of subject i. Then the observations consist of $\{C_i, Z_i, T_{ik}; 1 \le i \le n, 1 \le k \le m_i\}$ for $m_i > 1$ and $\{C_i, Z_i; 1 \le i \le n\}$ for $m_i = 0$. (Note that $T_{i,m_i+1} > C_i$)

2. gap times: let T^*_{ik}, $i = 1, \cdots, n$, $k = 1, \cdots, m_i$, be the time from $(k-1)$th recurrence to the kth recurrence of subject i. Then the data can be represented equivalently as $\{C_i, Z_i, T^*_{ik}; 1 \le i \le n, 1 \le k \le m_i\}$ for $m_i > 1$ and $\{C_i, Z_i; 1 \le i \le n\}$ for $m_i = 0$, where $\sum_{k=1}^{m_i+1} T^*_{ik} > C_i$.

3. counting process notation: let $N_i(t) = \sum_{k=1}^{\infty} I(T_{ik} \le t)$ denote the counting process recording the number of events occurred on subject i up to time t. Then the observations contain $\{C_i, Z_i(t), N_i(t) : 0 \le t \le C_i, i = 1, \cdots, n\}$.

- for clustered events data: let T_{ik} denote the event time of the kth subject in the ith cluster, $i = 1, \cdots, n$, $k = 1, \cdots, n_i$, where n_i is the total number of subjects in cluster i. In addition, let C_{ik} and Z_{ik} be the censoring and covariate variables of the kth subject in the ith cluster. Then the observations consist of $\{\tilde{T}_{ik}, \delta_{ik}, Z_{ik}; 1 \le i \le n, 1 \le k \le m_i\}$, where $\tilde{T}_{ik} = \min(T_{ik}, C_{ik})$ and $\delta_{ik} = I(T_{ik} \le C_{ik})$. (Question: what are the corresponding counting process and at-risk process for an individual study subject?)

Statistical methods for recurrent events data

1. Time-to-first-event analysis

The method only consider first event times and ignore all other succeeding event times. In other words, we only consider the observations: $\tilde{T}_{i1} = \min(T_{i1}, C_i)$ and $\delta_{i1} = I(T_{i1} \le C_i)$, $i = 1, \cdots, n$, as in the univariate survival data analysis we discussed before. For example, let's consider a two sample problem, where $Z_i = 1/0$ denoting the treatment or control group. Then we can compare the survival distributions of time-to-first-event of the two groups using the log-rank test statistics. That is:

$$\begin{aligned} U_f &= \sum_x \left\{ dN_1^f(x) - \frac{Y_1^f(x) \times dN^f(x)}{Y^f(x)} \right\} \\ &= \sum_x dN^f(x) \left[z_{I(x)} - \bar{z}^f(x, 0) \right], \end{aligned}$$

where $N^f(x) = \sum_{i=1}^n \delta_i I(\tilde{T}_{i1} \le x)$, $N_1^f(x) = \sum_{i=1}^n Z_i \delta_i I(\tilde{T}_{i1} \le x)$, $Y^f(x) = \sum_{i=1}^n I(\tilde{T}_{i1} \ge x)$ and $Y_1^f(x) = \sum_{i=1}^n Z_i I(\tilde{T}_{i1} \ge x)$. (Question: What is the formulation for $\bar{z}^f(x, 0)$? And what is the

relationship between $dN_1^f(x)$ and $dN^f(x)$?)

• Advantages: it is simple and nonparametric in nature.

• Disadvantages: ???

• How to extend the log-rank test to include all the recurrent events?

Actually, it is not difficult. Define the counting and at-risk processes for the recurrent events as follows: $N^R(x) = \sum_{i=1}^n N_i(x \wedge C_i)$, $N_1^R(x) = \sum_{i=1}^n Z_i N_i(x \wedge C_i)$, $Y^R(x) = \sum_{i=1}^n I(C_i \geq x)$ and $Y_1^R(x) = \sum_{i=1}^n Z_i I(C_i \geq x)$. The a log-rank type test statistics can be defined as

$$\begin{aligned} U_R &= \sum_x \left\{ dN_1^R(x) - \frac{Y_1^R(x) \times dN^R(x)}{Y^R(x)} \right\} \\ &= \sum_x dN^R(x) \left[z_{I(x)} - \bar{z}^R(x,0) \right], \end{aligned}$$

where $\bar{z}^R(x,0) = Y_1^R(x)/Y^R(x)$.

• Does the test U_R based on all the recurrence events always have bigger efficiency (i.e bigger power to detect the difference in survival distributions with the same sample size) that the test U_f based only on the time-to-first-event times?

2. Cox type multiplicative intensity model (see the seminar paper by Andersen and Gill (1982), Annals of Statistics, 1100-1120)

Model assumption: assume that $N_i(t)$ is a non-homogeneous poisson process with the intensity function $\lambda_0(t)\exp\{\beta' Z_i(t)\}$, i.e.

$$E\{dN_i(t)|N_i(u), Z_i(u), 0 \leq u < t\} = \lambda_0(t)\exp\{\beta' Z_i(t)\}dt.$$

Based on this model, a partial-likelihood type score can be constructed as:

$$U_R(\beta) = \sum_{\{\text{all grid pts } u\}} dN^R(u) \left[z_{I(u)}(u) - \frac{\sum_{l=1}^n z_l(u)\exp(\beta' z_l(u)) Y_l^R(u)}{\sum_{l=1}^n \exp(\beta' z_l(u)) Y_l^R(u)} \right],$$

then we solve $U^R(\beta) = 0$ for an estimator of β.

To fit the above model using R, we still use the coxph function. However, the data must be formatted in an appropriate way.

Inference procedures: By the martingale CLT, we have:

$$U_R(\beta_0) \overset{a}{\sim} \mathrm{N}(0, J_R(\beta_0)),$$

where $J_R(\beta)$ is the negative second derivative of the log-partial likelihood, i.e.

$$J_R(\beta) = \sum_u dN^R(u) \left[\frac{\sum_{l=1}^n z_l^2(u)\exp(\beta' z_l(u))Y_l^R(u)}{\sum_{l=1}^n \exp(\beta' z_l(u))Y_l^R(u)} - \left(\frac{\sum_{l=1}^n z_l(u)\exp(\beta' z_l(u))Y_l^R(u)}{\sum_{l=1}^n \exp(\beta' z_l(u))Y_l^R(u)} \right)^2 \right].$$

In addition, treating $J_R(\beta_0)$ as a constant, we get the approximate distribution of $(\hat{\beta} - \beta_0)$

$$(\hat{\beta} - \beta_0) \overset{a}{\sim} \mathrm{N}(0, J_R^{-1}(\beta_0)).$$

• Restrictions of the Cox type multiplicative intensity model: nonhomogeneous poisson assumption essentially assumes that given covariates, the gap times from the same subject are independent (i.e. independent increment property), which is often too restrictive in practice.

3. Proportional mean/rate model (a marginal model)

The proportional mean model relaxes the nonhomogeneous poisson assumption of the Cox type multiplicative intensity model, and only assume that the marginal expectation of $dN_i(t)$ given covariates has the form

$$E\{dN_i(t)|Z_i(u), 0 \le u < t\} = \lambda_0(t)\exp\{\beta' Z_i(t)\}dt.$$

The estimation procedure is the same as for the Cox type multiplicative intensity model. However, the inference procedure is different. Now we have, by the modern empirical process theory:

$$U_R(\beta_0) \overset{a}{\sim} \mathrm{N}(0, \Sigma_R(\beta_0)),$$

and

$$(\hat{\beta} - \beta_0) \overset{a}{\sim} \mathrm{N}(0, J_R^{-1}(\beta_0)\Sigma(\beta_0)J_R^{-1}(\beta_0)),$$

where in general $\Sigma_R(\beta_0) \neq J_R(\beta_0)$.

4. Frailty model (Andersen, Borgan, Gill, and Keiding, 2000)

Model assumption: assume that given a positive latent variable or frailty η_i, $N_i(t)$ is a non-homogeneous poisson process with the intensity function $\eta_i \lambda_0(t) \exp\{\beta' Z_i(t)\}$, where $E(\eta_i) = 1$ and $Var(\eta_i) = \sigma^2$.

Here η_i is a subject-specific random effect, which is unobservable. The variance σ^2 of η_i measures the magnitude of within subject correlation. When $\sigma^2 = 0$, i.e. $\eta_i \equiv 1$, the frailty model becomes the Cox type multiplicative intensity model. However, when $\sigma^2 > 0$, the frailty model allows correlations between gap times of the same subject even after adjusting for covariates. Note that, marginally, i.e. given the covariates $Z_i(t)$ only, $N_i(t)$ is no longer a nonhomogeneous poisson process.

Statistical methods for clustered events data

Here for simplicity of the representation, we assume that each cluster only contains two subjects, like the diabetic retinopathy study or the Australian twin study.

1. Frailty model

Given the cluster-specific frailty η_i and covariates Z_{ik}, we consider the following proportional hazards model for the event time T_{ik}, $i = 1, \cdots, n$, $k = 1, 2$,

$$\lambda(t|\eta_i, Z_{ik}) = \eta_i \exp(\beta' Z_{ik}) \lambda_0(t),$$

where $\lambda_0(t)$ is a completely unspecified baseline hazard function. Note that after integrating the random effect η_i out, marginally T_{ik} is no longer from the proportional hazards model (why?). And could you write down the joint likelihood function of the data from the ith cluster?

2. Marginal model

The model directly assumes that given covariates Z_{ik}, T_{ik} marginally follows the proportional

hazards model $\lambda(t|Z_{ik}) = \exp(\beta' Z_{ik})\lambda_0(t)$, while the dependence structure between the event times within the same cluster is left completely unspecified.