Survival Data Analysis (BIOS:7210)
Breheny

Assignment 2
Due: Tuesday, September 17

1. In class, we looked at the hazard function of the exponential distribution. Consider the hazard function for the gamma distribution.

   (a) Choose a few different shape parameters that represent the variety of hazard function curves that are possible with the gamma distribution and plot the resulting hazard.

   (b) Describe a specific situation in which the gamma distribution would be a reasonable parametric model for a time to an event, as well as a specific situation in which it would not be reasonable.

2. In each scenario below, write the likelihood for the data assuming independent censoring. Let $\lambda(t)$, $f(t)$, $F(t)$, and $S(t)$ denote the hazard, density, distribution, and survival functions, respectively, of the time to the event of interest.

   For each problem, write the likelihood in terms of dates that would need to be recorded as part of the study (i.e., part of this problem is thinking about the logistics of the data collection). For example, suppose the problem was:

   > Patients newly diagnosed with lung cancer are followed; the study continues until all patients have died. The study wishes to estimate the time from diagnosis until death.

   An acceptable answer would be:

   > Letting a_i denote the date of diagnosis and b_i denote the date of death for individual i,
   >
   > $$L = \prod_{i=1}^{n} f(b_i - a_i),$$
   >
   > where n denotes the sample size.

   For some problems, you may require additional notation in order to write the likelihood; introduce what you need, but clearly define it.

   (a) Patients newly diagnosed with Hodgkin's lymphoma are followed for 10 years in order to study the time from diagnosis until death from cancer. Not all patients are observed to die during the study period, for a variety of reasons (death from other causes, withdrawal from the study, still alive at end of study, etc.).

   (b) A study involving elementary school children is investigating the ages at which children break their arm for the first time. When the children enter first grade, they are given an X-ray to determine if they have previously broken their arm prior to entering school. If at any point during their time in elementary school, they break their arm, this is recorded. The arms of some children, of course, will remain unbroken upon their graduation from elementary school.

(c) Workers at a chemical plant undergo a medical exam every 2 years looking for possible signs of oral cancer. No workers had oral cancer at the beginning of their employment. The study wishes to estimate the distribution of time until oral cancer since first employment at the plant.

(d) A study of human mortality (age at death) is carried out among the residents of a retirement center. All individuals in the retirement center are followed until death, but of course not everyone in the population will go to this retirement center.

3. Show that the random censoring likelihood we derived in class,

$$L(\theta) = \prod_i f(t_i|x_i,\theta)^{d_i} S(t_i|x_i,\theta)^{1-d_i},$$

is equivalent to the independent censoring formulation of the likelihood:

$$L(\theta) = \left\{\prod_i \lambda(t_i|\theta,x_i)^{d_i}\right\} \exp\left\{-\int_0^\infty \sum_{k\in R(u)} \lambda(u|\theta,x_k)\,du\right\},$$

where $R(u)$ is the risk set at time u, consisting of all the individuals still alive and uncensored at time u.

4. Suppose the survival time $T_i \overset{\perp\!\!\!\perp}{\sim} \text{Exp}(1)$ for $i = 1,\ldots,n$, but we only sample M subjects with $T < 1$ (nothing is known about subjects with $T \geq 1$, including how many of them there are). Note that M is a random variable in this formulation. We saw in class that with only a small number of observations $n = 5$, we were unable to estimate λ accurately (the likelihood was extremely wide). For three simulated data sets, with sample sizes of $n = 100, n = 1,000$, and $n = 10,000$, construct and plot the likelihood as we did in class. Comment on whether the likelihood appears consistent (i.e., it becomes more and more concentrated on the true value of λ) as well as its rate of convergence (are we rapidly learning what the true value of λ is, or is it taking a really long time?).

5. Simulate $n = 500$ observations of true failure times from a Gamma($\alpha = 1/2, \lambda = 1$) distribution. Simulate n censoring times from an Exp(1) distribution. As usual, the observed data consist of the times on study as well as the failure/censoring indicator.

(a) Treating $\alpha = 1/2$ as known, plot the likelihood $L(\lambda)$.

(b) Treating $\lambda = 1$ as known, plot the likelihood $L(\alpha)$.