An introduction to survival models

Xiang Ao

January 12, 2009

1 Survival function

Survival analysis is also called time to event analysis or event history analysis. We use survival models to analyze data that have the following characteristics:

1. the dependent variable is the waiting time till the occurrence of some event;
2. some observations are censored; that is, some units are not in the data set anymore for some reason before an event happens. For some units we observe the event and the exact waiting time; for others it has not occurred, and all we know is that the waiting time exceeds the observation time. However, we believe if they stayed in the study, sooner or later an event would happen.
3. we are trying to explain (predict) the occurrence of an event with some predictors (explanatory variables).

Let T denote the time variable, a nonnegative, continuous random variable with PDF $f(t)$ and CDF $F(t)$, where t is a realization of T. The survival function is defined by

$$S(t) = Pr(T > t) = 1 - F(t) \tag{1}$$

The graph of $S(t)$ against t is known as the survival curve. If there is no censoring, then the survival function is easy to estimate by the ratio of number of units survives at t vs. total number of units at risk at t. The most common method to estimate the survival function with a survival data set containing censored observations is the Kaplan-Meier method. The basic idea is to use the product of a series of conditional probabilities.

First sort the event times from the smallest to the largest:

$$t_{(1)} \leq t_{(2)} \leq ... \leq t_{(r)} \tag{2}$$

Then the survival curve is estimated by:

$$\hat{S}(t) = \prod_{j|t(j)\leq t} (1 - \frac{d_j}{r_j}), \tag{3}$$

where r_j is the number of units at risk at $t_{(j)}$ and d_j is the number of "failure" or events at $t_{(t)}$. Note that units censored at $t_{(j)}$ are included in r_j.

The variance of this estimator can be estimated by:

$$\hat{var}(S(t)) = (\hat{S}(t))^2 \sum_{j|t(j)\leq t} \frac{d_j}{r_j(r_j - d_j)}. \tag{4}$$

2 Hazard function

We are often interested in which periods have the highest or lowest change of "death" or "failure" or other events. That is, we may be interested in the probability that a state ends between t and $t+\Delta t$ conditional on having reached t in the first place.

The probability is

$$\Pr(t < T \leq t + \Delta t | T \geq t) = \frac{F(t + \Delta t) - F(t)}{S(t)} \tag{5}$$

The hazard function is defined by

$$h(t) = \lim_{\Delta t \to 0} \Pr(t < T \leq t + \Delta t | T \geq t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)} \tag{6}$$

One of the simplest functional forms is the exponential distribution

$$f(t, \theta) = \theta e^{-\theta t} \tag{7}$$

Therefore, the hazard function is

$$h(t) = \theta \tag{8}$$

A much more flexible functional form is Weibull

$$f(t, \theta, \alpha) = 1 - \exp(-(\theta t)^{\alpha}). \tag{9}$$

Hazard function can be shown to be

$$h(t) = \alpha \theta^{\alpha} t^{\alpha - 1} \tag{10}$$

However, Weibull distribution does not allow for the possibility that the hazard may first increase then decrease over time. Log-normal distribution does allow this.

Figure 1: Various hazard functions from Davidson and MacKinnon

3 Log-likelihood function

Suppose we have n units with a survival function $S(t)$, density $f(t)$, and hazard $h(t)$. If for unit i, the event happens at t_i, it's contribution to the likelihood function is the density at that point:

$$L_i = f(t_i) = S(t_i)h(t_i) \tag{11}$$

If no event at t_i, all we know is that the lifetime exceeds t_i. The probability is

$$L_i = S(t_i), \tag{12}$$

which is the contribution to the likelihood function from a censored observation.

Therefore, if U denotes the set of uncensored observations, the log-likelihood function for the entire sample can be written as

$$\ell(t, \theta) = \sum_{i \in U} \log h(t_i|\mathbf{X}_i, \theta) + \sum_{i=1}^{n} \log S(t_i|\mathbf{X}_i, \theta) \tag{13}$$

4 Proportional Hazard Models

4.1 Model and likelihood

One class of models that is quite widely used is the proportional hazard models (based on Cox (1972)).

$$h(t|\mathbf{x}_i) = h_0(t) \exp(\mathbf{x}_i'\beta), \tag{14}$$

In this model $h_0(t)$ is called baseline hazard rate; it describes the risk for units with $\mathbf{x_i} = \mathbf{0}$. $\exp(\mathbf{x}_i'\beta)$ represents the relative risk, a proportionate increase or decrease in risk, due to $\mathbf{x_i}$. The interpretation of β is that $\exp(\beta_i)$ gives the relative risk change associated with an increase of one unit in x_i, all other explanatory variables remaining constant.

β is estimated by maximizing a partial likelihood function. It is partial likelihood function since the baseline hazard is factored out. Suppose one event happens at time t_j. Conditional on this event the probability that case i dies (meaning this event happened to case i, not other cases in the risk set) is

$$L_i = \frac{h_0(t) \exp(x_i'\beta)}{\sum_l I(T_l \geq t) h_0(t) \exp(x_l'\beta)} = \frac{\exp(x_i'\beta)}{\sum_l I(T_l \geq t) \exp(x_l'\beta)} \tag{15}$$

We can see here that Cox's model is for continuous time survival data. It is assumed there is no ties at time t_j (in continuous time is not possible for two events to happen at the same time). In reality, we observe ties of event

time, because in many situations, we observe grouped data. For example, we observe case m and n have events at the same time j. In that case, in the denominator, whether we include case m in the calculation of the likelihood of case n is ambiguous. In that case, special treatment is needed. A standard approximation (Breslow and Peto) is to let

$$L_{m\in D(t_j)} \simeq \frac{\prod_{m\in D(t_j)} \exp(x'_m\beta)}{[\sum_{l\in R(t_j)} \exp(x'_l\beta)]^{d_j}} \tag{16}$$

where $D(t_j)$ is the set of cases that die at time t_j and d_j denotes the number of cases that die at time t_j. This approximation works well with small number of ties relative to total number of cases at risk.

One important feature (disadvantage?) for the proportional hazard model is that for all units, the effect is the same at all time t. To see this:

$$\frac{h_1(t)}{h_2(t)} = \frac{h_0(t)\exp(\mathbf{x_1}'_i\beta)}{h_0(t)\exp(\mathbf{x_2}'_i\beta)} = \frac{\exp(\mathbf{x_1}'_i\beta)}{\exp(\mathbf{x_2}'_i\beta)}, \tag{17}$$

which does not depend on t.

One advantage of Cox's model is that it is considered as a semi-parametric model since it does not have to specify the distribution of the survival time. The estimation process relies only on the order in which events occur, not the exact time they occur. Parameter estimates are derived assuming continuous survival times.

4.2 Interpretation

In equation **??**, if we would like to calculate the partial effect of x_j on h, we have:

$$\partial h(t|\mathbf{x}_i)/\partial x_j = h_0(t)\exp(\mathbf{x}'_i\beta)\beta_j = \beta_j h(t|\mathbf{x}_i), \tag{18}$$

If we do it the other way, plug in $x_j + 1$ will generate

$$h_0(t)\exp(\mathbf{x}'_i\beta + \beta_j) = \exp(\beta_j)h(t|\mathbf{x}'_i\beta) \tag{19}$$

Therefore, one unit change in x_j will have a change of $1 - \exp(\beta_j)$ times the original hazard.

5 Discrete-time Survival Models

Cox (1972) proposed a discrete-time model by modeling the odds of dying at time t_j given survival up to that point.

$$\frac{h(t_j|\mathbf{x}_i)}{1-h(t_j|\mathbf{x}_i)} = \frac{h_0(t_j|\mathbf{x}_i)}{1-h_0(t_j|\mathbf{x}_i)}\exp(\mathbf{x}'_i\beta), \tag{20}$$

which is similar to Cox's proportion hazard model for continuous time. If we take logs on both sides, we have

$$\text{logit}(h(t_j|\mathbf{x}_i)) = \alpha_j + \mathbf{x}_i'\beta, \tag{21}$$

This looks very similar to a logit model. In fact, we can fit a discrete-time proportional hazard model by running a logistic regression on a set of pseudo observations generated by "filling in" the observations between the start time to the event time or time at the end of the study period (censored). The outcome of this process is 0 unless there is an event, then it turns 1. It is treated as independent Bernoulli observations with probability given by hazard h_{ij} for individual i at time t_j. The likelihood function for the discrete-time survival model coincide with the binomial likelihood of independent binomial process ([**?**] has more details).

6 Do it in Stata

To do a discrete-time survival model in Stata, your data set needs to be set up as by unit-time. For example, observations by company month. The event (or failure) variable needs to be set to zero until the event happens. If no event until the end of the study period, then it is censored. A data set by unit-time can be analyzed by logit model, optionally with clustered standard error. To set up a structure like this, user-written programs called *dthaz* and *prsnperd* (means person-period) can be used.

To do a continuous-time survival model in Stata, your data can be either set up as discrete-time case (in which case you can use a time-varying explanatory variable) or a single-record-per-person data set. Stata has build-in command *stset* to set up a survival analysis structure. Then you can run *stcox* for Cox's proportional hazard model.