STAT 425: Introduction to Nonparametric Statistics **Winter 2018**

Lecture 5: Survival Analysis

Instructor: Yen-Chi Chen

Note: in this lecture, we will use the notations $T_1, \cdots, T_n$ as the response variable and all these random variables are positive. These random variables will be called event time or death time. They often refer to certain 'time' characteristics of each individual, e.g., the time that the individual is dead/gets a disease.

5.1 Survival Function

We assume that our data consists of IID random variables $T_1, \cdots, T_n \sim F$. The **survival function** $S(t)$ of this population is defined as

$$S(t) = P(T_1 > t) = 1 - F(t).$$

Namely, it is just one minus the corresponding CDF. Although this definition is extremely simple and seems to be very trivial from the CDF, later we will see that it turns out to be an elegant tool of modeling and interpreting the data.

In medical research, the quantity T_i often refers to certain time characteristic of individual i. For instance, the variable T may refer to the age that the individual i passes away. Then the survival function $S(t)$ can be interpreted as *the chance that an individual is still alive after age t*. If $S(60) = 0.8$, it means that there are 80% of the individuals in the population who will still be alive at the age 60. Namely, $S(t)$ is the probability that an individual will survive past time t.

Here are some basic properties about $S(t)$:

- $S(0) = 1$ and $S(\infty) = 0$.
- $S(t)$ is a non-increasing function.

A quantity that is often used along with the survival function is the hazard function. The **hazard function** is

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T_1 \leq t + \Delta t | T_1 > t)}{\Delta t} = \frac{p(t)}{S(t)},$$

where $p(t) = \frac{d}{dt}F(t)$ is the PDF of random variable T_1. Note that you can also write the hazard function as

$$h(t) = -\frac{\partial \log S(t)}{\partial t}.$$

How can we interpret the hazard function? The hazard function describes the 'intensity of death' at the time t given that the individual has already survived past time t.

There is another quantity that is also common in survival analysis, the cumulative hazard function. The **cumulative hazard function** is

$$H(t) = \int_0^t h(s)ds.$$

You can interpret $H(t)$ as the cumulative amount of hazard up to time t. The cumulative hazard function and survival function as linked as follows:

$$H(t) = -\log S(t), \quad S(t) = e^{-H(t)} = e^{-\int_0^t h(s)ds}.$$

Example 1. What is the survival function and hazard function of an exponential R.V.? Let $T_1 \sim \mathsf{Exp}(\lambda)$. Then

$$p(t) = \lambda e^{-\lambda t}, \quad F(t) = 1 - e^{-\lambda t} \text{ for } t \geq 0$$

Thus,

$$S(t) = e^{-\lambda t}$$

and

$$h(t) = \lambda, \quad H(t) = \lambda t.$$

Namely, in an exponential distribution, the hazard function is a constant and the cumulative hazard is just a linear function of time.

Example 2 (Weibull distribution). The Weibull distribution is a distribution with two parameters, λ and k, and it is a distribution for positive random variable. Its PDF is

$$p(t) = \lambda k \cdot (\lambda t)^{k-1} \cdot e^{-(\lambda t)^k}, t \geq 0.$$

When $k = 1$, it reduces to the exponential distribution. Its CDF and survival function are

$$F(t) = 1 - e^{-(\lambda t)^k}, \quad S(t) = e^{-(\lambda t)^k}.$$

And the hazard function and cumulative hazard function are

$$h(t) = \lambda k \cdot (\lambda t)^k, \quad H(t) = (\lambda t)^k.$$

5.1.1 Estimating the Survival Function: Simple Method

How do we estimate the survival function? There are three methods. The first method is a parametric approach. This method assumes a parametric model (e.g., exponential distribution) of the data and we estimate the parameter first then form the estimator of the survival function. A second approach is to compute the EDF first and then converted it to an estimator of the survival function. The last approach is a powerful nonparametric method called the Kaplan-Meier estimator and we will discuss it in the next section.

Parametric Approach. Assume that we model the distribution as an exponential distribution with unknown parameter λ. An estimator of λ is (you can check HW01 to see why this is an estimator)

$$\widehat{\lambda} = \frac{1}{\bar{T}_n} = \frac{n}{\sum_{i=1}^n T_i}.$$

Then we estimate the survival function using

$$\widehat{S}_1(t) = \widehat{\lambda} e^{-\widehat{\lambda} t} = \frac{e^{-\frac{t}{\bar{T}_n}}}{\bar{T}_n}, \quad t \geq 0.$$

EDF Approach. Recall that the EDF $\widehat{F}(t)$ will be

$$\widehat{F}(t) = \frac{1}{n} \sum_{i=1}^n I(T_i \leq t).$$

Then the survival function can be estimated by

$$\widehat{S}_2(t) = 1 - \widehat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} I(T_i > t).$$

5.1.2 Kaplan-Meier estimator

Let $t_1 < t_2 < \cdots < t_m$ be the time point where the observations $T_1, \cdots, T_n$ actually take values.

To see how the estimator is constructed, we do the following analysis. We partition the time axis into disjoint segments:

$$B_0 = [0, t_1), B_1 = [t_1, t_2), \cdots, B_{m-1} = [t_{m-1}, t_m), B_m = [t_m, \infty).$$

Then we define

$$N_\ell = \text{ number of individuals alive at (event happens after) the beginning of } B_\ell = \sum_{i=1}^{n} I(T_i \geq t_\ell)$$

and

$$D_\ell = \text{ number of individuals die (event happens at) in } B_\ell = \sum_{i=1}^{n} I(T_i \in B_\ell).$$

Now we have converted $T_1, \cdots, T_n$ to $(N_0, D_0), \cdots, (N_m, D_m)$. Formally, N_ℓ should be defined as the number of individuals *at risk* at the beginning of B_ℓ. Later we will explain what does the *at risk* means.

The **Kaplan-Meier (KM) estimator** estimates $S(t)$ using

$$\widehat{S}_{KM}(t) = \prod_{\ell: t_\ell \leq t} \left(1 - \frac{D_\ell}{N_\ell}\right).$$

What is the intuition of the KM estimator? We now consider t in different time segments and see if we can gain some intuitions. Recall that the survival function

$$S(t) = P(T > t) = \text{ Probability of surviving past time } t.$$

For $t \in B_0 = [0, t_1)$, there is no event happens within this interval so $\widehat{S}_{KM}(t) = 1$.

For $t \in B_1 = [t_1, t_2)$, the survival function

$S(t) = P(T > t) = P(\text{survives past time } t) = P(\text{survives in } [0, t_1) \text{ and in } [t_1, t)) = P(\text{survives in } B_0 \text{ and in } B_1)$.

Now recall that for two events A and B, $P(A \text{ and } B) = P(A)P(B|A)$. Thus,

$$S(t) = P(\text{survives in } B_0 \text{ and in } B_1) = P(\text{survives in } B_0)P(\text{survives in } B_1|\text{survives in } B_0).$$

The probability $P(\text{survives in } B_1|\text{survives in } B_0)$ can be estimated using

$$\widehat{P}(\text{survives in } B_1|\text{survives in } B_0) = \frac{N_1 - D_1}{N_1} = 1 - \frac{D_1}{N_1}$$

and because no event occurs in B_0, $P(\text{survives in } B_0) = 1$. Thus,

$$\widehat{S}_{KM}(t) = 1 \times \left(1 - \frac{D_1}{N_1}\right).$$

Now for the next time segment B_2, we apply the same intuition. Namely, for $t \in B_2$,

$$S(t) = P(\text{survives in } B_0)P(\text{survives in } B_1|\text{survives in } B_0)P(\text{survives in } B_2|\text{survives in } B_1),$$

where we can estimate $P(\text{survives in } B_2|\text{survives in } B_1)$ via

$$\widehat{P}(\text{survives in } B_2|\text{survives in } B_1) = 1 - \frac{D_2}{N_2},$$

which leads to

$$\widehat{S}_{KM}(t) = 1 \times \left(1 - \frac{D_1}{N_1}\right) \times \left(1 - \frac{D_2}{N_2}\right).$$

For the other segments, we can apply the same procedure to obtain the estimator. This gives you the intuition of how the KM estimator is constructed. This derivation can also be seen in http://pages.stat.wisc.edu/~ifischer/Intro_Stat/Lecture_Notes/8_-_Survival_Analysis/8.2_-_Kaplan-Meier_Formula.pdf.

Note that when we observe every individual's event time (namely, there is no censoring – a mechanism we will discuss later), the KM estimator and the EDF approach are the same.

5.1.3 Nelson-Aalen estimator

Nelson-Aalen (NA) estimator is another powerful estimator of the survival function. It not only estimates the survival function but also provides an estimate of the cumulative hazard. Actually, NA estimator first estimate the cumulative hazard function and then convert it into an estimate of the survival function using the relation $S(t) = e^{-H(t)}$. Here is an intuition about how this estimator is constructed.

Recall that the KM estimator uses

$$\widehat{S}_{KM}(t) = \prod_{\ell:t_\ell \le t} \left(1 - \frac{D_\ell}{N_\ell}\right).$$

as an estimate of $S(t)$. When D_ℓ is much smaller than N_ℓ, we have

$$e^{-\frac{D_\ell}{N_\ell}} \approx 1 - \frac{D_\ell}{N_\ell}.$$

Therefore,

$$\begin{aligned}
\widehat{H}_{KM}(t) &= -\log \widehat{S}_{KM}(t) \\
&= -\log \prod_{\ell:t_\ell \le t} \left(1 - \frac{D_\ell}{N_\ell}\right) \\
&= -\sum_{\ell:t_\ell \le t} \log \left(1 - \frac{D_\ell}{N_\ell}\right) \\
&\approx -\sum_{\ell:t_\ell \le t} \log e^{-\frac{D_\ell}{N_\ell}} \\
&= \sum_{\ell:t_\ell \le t} \frac{D_\ell}{N_\ell}.
\end{aligned}$$

Using the above derivation, the NA estimator estimates the cumulative hazard function by

$$\widehat{H}_{NA}(t) = \sum_{\ell:t_\ell \le t} \frac{D_\ell}{N_\ell}$$

and then estimate the survival function as

$$\widehat{S}_{NA}(t) = e^{-\widehat{H}_{NA}(t)} = e^{-\sum_{\ell:t_\ell \le t} \frac{D_\ell}{N_\ell}} = \exp\left(-\sum_{\ell:t_\ell \le t} \frac{D_\ell}{N_\ell}\right).$$

The theoretical analysis of the KM and NA estimators (such as the expectation and variance) involve some non-trivial algebra. If you are interested in, I would recommend the following lecture note http://www4.stat.ncsu.edu/~dzhang2/st745/chap2.pdf.

5.2 Censoring

However, in reality, our data may not be so nice. We may not be able to observe the actual event time T_i because of many complications. For instance, in a medical research, individuals may leave the study (called dropout) so we only observe their leaving time instead of the actual death time. The phenomena that we sometimes cannot observe the actual time but a 'censoring time' is called **censoring** in Statistics.

To model this process, we often need to introduce two other variables: Y and C. The T is the actual event time of interest and C is the censoring time that is competing with T and Y is the actual observing time.

In most cases, we will consider the **right-censoring** problem where the three variables are related by

$$Y = \min\{T, C\}.$$

We will assume that T and C are independent. Note that if what we observe is $Y = \max\{T, C\}$, this problem is called a left-censoring problem. Moreover, we not only observe Y, we also know if this Y comes from the event time or censoring time. Namely, we have one extra variable δ such that $\delta = I(T < C)$.

When we only observe $(Y_1, \delta_1), \cdots, (Y_n, \delta_n)$ instead of $T_1, \cdots, T_n$, how can we infer the survival function T_1? This is the central question to many biostatistical research.

Because we have several R.V.s now, we will add subscript to denote the functions associated to each random variable. Namely, F_T, S_T, h_T, H_T are the CDF, survival function, hazard function, and cumulative hazard function of random variable T and F_C, S_C, h_C, H_C are those of random variable C and F_Y, S_Y, h_Y, H_Y are those of random variable Y.

Here are some relations among these functions.

- $S_Y(t) = P(Y > t) = P(\min\{T, C\} > t) = P(T > t)P(C > t) = S_T(t)S_C(t)$.
  Namely, the survival function of Y is the product of the other two survival functions.
- $F_Y(t) = 1 - (1 - F_T(t))(1 - F_C(t)) = F_T(t) + F_C(t) - F_T(t)F_C(t)$.
- $p_Y(t) = p_T(t) + p_C(t) - p_T(t)F_C(t) - p_C(t)F_T(t) = p_T(t)S_C(t) + p_C(t)S_T(t)$.
  The PDF of Y is the sum of the weighted PDF of the other two and the weight is the survival function.
- $h_Y(t) = h_T(t) + h_C(t)$.
  Namely, the hazard function of Y is the summation of the other two.
- $H_Y(t) = H_T(t) + H_C(t)$.
  Similarly, the cumulative hazard is also the sum of the other two.

Note that δ is just a Bernoulli random variable with probability being 1 as $P(T < C)$.

5.2.1 Estimating the Survival Function in Censoring

When there is censoring, the EDF approach no longer works. However, the KM and NA estimators are still valid. Essentially, the estimator is the same but we need to modify a little bit about N_ℓ and D_ℓ. As we have mentioned, formally, N_ℓ should be defined as

$$N_\ell = \text{ number of individuals at risk at the beginning of } B_\ell.$$

What does the phrase *at risk* means? It refers to as being alive *and* not censored so it can be modified by replacing T_i with Y_i. Thus,

$$N_\ell = \sum_{i=1}^{n} I(Y_i \geq t_\ell).$$

For the quantity D_ℓ, it is still the number of events in the interval B_ℓ but we need to modify it by the number of *observed* events in the interval. Therefore,

$$D_\ell = \sum_{i=1}^{n} I(Y_i \in B_\ell, \delta_i = 1).$$

Using these two modifications, the KM estimator and NA estimator are

$$\widehat{S}_{KM}(t) = \prod_{\ell: t_\ell \leq t} \left(1 - \frac{D_\ell}{N_\ell}\right)$$

$$\widehat{S}_{NA}(t) = \exp\left(-\sum_{\ell: t_\ell \leq t} \frac{D_\ell}{N_\ell}\right).$$

Note that parametric models may still be applicable during the censoring case and the estimator is often done using a maximum likelihood approach, which is beyond the scope of this course so we will not cover it here. Here is a lecture note about this topic: http://www4.stat.ncsu.edu/~dzhang2/st745/chap3.pdf

5.3 Cox Model

In reality, we often not only observe the event time for an individual but also have access to other covariates of this individual. We often are interested in understanding how these covariates affect the survival function of the event.

For instance, in a cancer study, we may have each individual's age when they got cancer (the event time T) and this individual's gender, BMI, smoking habit, and education level. The other variables are the covariates in this study. Health scientists are often interested in how these covariates change the survival function. Let X denotes the covariates. A parameter of interest will be the survival function of T given X. Namely, it is the conditional survival function

$$S(t|x) = P(T > t|X = x).$$

For instance, we may be interested in

$$S(\text{Age} = t|(\mathsf{gender}, \mathsf{BMI}, \mathsf{smokinghabit}, \mathsf{educationlevel}) = (\mathsf{male}, 20, \mathsf{neversmoke}, \mathsf{college})).$$

We can then define the conditional hazard function and conditional cumulative hazard function as

$$h(t|x) = -\frac{\partial \log S(t|x)}{\partial t}, \quad H(t|x) = -\log S(t|x).$$

The **Cox (proportional hazard) model** is one of the most popular model combining the covariates and the survival function. It starts with modeling the hazard function $h(t|X = x)$:

$$h(t|X = x) = h_0(t) \exp(x^T \beta),$$

where β is the vector of coefficients of each covariate. The function $h_0(t)$ is called the baseline hazard function. Namely, the Cox model assumes that the covariates have a linear multiplication effect on the hazard function and the effect stays the same across time.

This implies the conditional hazard function being

$$H(t|x) = \exp(x^T \beta) \int_0^t h_0(s) ds = \exp(x^T \beta) H_0(t),$$

where $H_0(t)$ is the baseline cumulative hazard function. This further yields the conditional survival function

$$S(t|x) = \exp(-H(t|x)) = \exp\left(-\exp(x^T \beta) H_0(t)\right) = \exp\left(-H_0(t)\right)^{\exp(x^T \beta)} = S_0(t)^{\exp(x^T \beta)},$$

where $S_0(t)$ is called the baseline survival function.

Why it is called a *proportional* hazard model? Here is an intuition about it. Consider two individuals with different covariates that one has $X = x_1$ and the other has $X = x_2$. The ratio of their hazard function

$$\frac{h(t|x_1)}{h(t|x_2)} = \frac{h_0(t) \exp(x_1^T \beta)}{h_0(t) \exp(x_2^T \beta)} = \frac{\exp(x_1^T \beta)}{\exp(x_2^T \beta)} = \exp((x_1 - x_2)^T \beta)$$

is a constant over time. Namely,

$$h(t|x_1) = \exp((x_1 - x_2)^T \beta) \times h(t|x_2) \propto h(t|x_2) \quad \forall t \geq 0.$$

Thus, their hazard is always proportional to each other regardless of the value of time t.

Estimation of the parameter β is often done by maximizing the *partial likelihood function*:

$$\widehat{L}_n(\beta) = \prod_{i=1}^{n} L_i(\beta),$$

where

$$L_i(\beta) = \frac{h(T_i|X_i)}{\sum_{j:T_j \geq T_i} h(T_j|X_j)} = \frac{\exp(X_i^T \beta)}{\sum_{j:T_j \geq T_i} \exp(X_j^T \beta)}.$$

Namely, our estimator

$$\widehat{\beta}_n = \mathsf{argmax}_\beta \ \widehat{L}_n(\beta).$$

This estimator turns out to be an unbiased estimator and has variance shrinking at rate $O(n^{-1})$ and has asymptotic normality under suitable condition. An interesting fact is that *we do not need to know the baseline hazard function $h_0(t)$ to estimate β!* (estimating $h_0(t)$ is not easy and the convergence rate is often slow; we will discuss a similar pattern in density estimation) The property that we can estimate parameter of interest without estimating the entire model is related to the topic *semi-parametric model*[1].

Note that the detailed analysis an derivation is beyond the scope of this course (you may learn it in a course called 'survival analysis'). If you want to learn more, I would recommend the following two lecture notes:

- http://www4.stat.ncsu.edu/~dzhang2/st745/chap6.pdf
- http://www.public.iastate.edu/~kkoehler/stat565/coxph.4page.pdf

[1] https://en.wikipedia.org/wiki/Semiparametric_model