Independent Censorship Model for Survival Data

Mai Zhou

We consider right censored survival data:

$$3+, 6, 2.2, 8+, 12, \cdots.$$

They are commonly recorded as two vectors (usually called *observed survival times* and *status*), instead of a 'plus':

$$\begin{aligned} \text{censored survival times:} \quad T &= (3, 6, 2.2, 8, 12, \cdots) \\ \text{censoring status:} \quad \delta &= (0, 1, 1, 0, 1, \cdots); \end{aligned}$$

with $(T_1, \delta_1) = (3, 0)$, $(T_2, \delta_2) = (6, 1), \cdots$, etc.

We assume the observed survival time T is the minimum of two (sometimes latent) values:

$$T_i = \min(X_i, C_i)$$

where X_i is the true lifetime under study if we were to follow the patient forever; and C_i is the (finite) follow-up time.

Sometimes, C_i is the competing risk disease (of all other causes), sometime it is the end of the study set by design.

The distribution of X is what we are trying to estimate, test or model.

The status or the '+' sign is just an indication of the observed times T is due the competing risk/end of study. So,

$$\delta = 1 \ \text{ if } \ T = X; \quad \text{ or } \quad \delta = 0 \ \text{ if } \ T = C$$

in short, $\delta = I[X \leq C]$.

We further assume the random variables X is independent of C.

For $i = 1, 2, \cdots, n$, assume X_i are independent and identically distributed lifetimes with CDF $F(t)$. Further we suppose C_i are also independent identically distributed censoring times with CDF $G(t)$. We suppose that the X_i's are independent of the C_i's. What we observe are $T_i = \min(X_i, C_i)$ and $\delta_i = I[X_i \leq C_i]$.

Our task is to estimate/test/model the distribution of X, assume the independent censorship model. The data we can use are the vector T's and δ's. (not X_i not C_i).

• Without the independent assumption, we may not be able to estimate the distribution of X alone since the distribution of X and C may not be separable or called

identifiable. But there are some relaxation using the concept of 'conditional independency'.

• dependent among X_i or C_i can be handled a bit easier. (α-mixing)

• identical distribution among X_i or C_i are less crucial but the estimator is estimating a sort of average.

Reference: Tsiatis, A. (1975) A Nonidentifiability Aspect of the Problem of Competing Risks Page 20 of 20-22.

Now we see a unique problem: In traditional statistical inference we observe x_i and want to estimate the distribution of x_i, (what we observe is from what we want to estimate.)

In survival analysis, we observe T_i, δ_i and want to estimate the distribution of X_i, (what we observe is *not* from what we want to estimate).

Therefore, some kind of model or assumption is a must.

Solution: write down the likelihood of sample T_i, δ_i, but in terms of distribution of X_i.