Likelihood for Censored Data

Mai Zhou

We consider right censored survival data:

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

They are commonly recorded as two vectors, instead of a 'plus':

$$T = (3, 6, 2.2, 8, 12, \cdots)$$

$$\delta = (0, 1, 1, 0, 1, \cdots);$$

with $(T_1, \delta_1) = (3, 0)$ etc.

Assume $(T_1, \delta_1), (T_2, \delta_2), \cdots$ are iid (2-dimensional) random vectors from a two dimensional distribution; with one component continuous (T) and one component discrete (δ, only two possible values).

The likelihood function is the probability of the observed sample under assumed model, be it parametric or nonparametric:

$$\prod_{i=1}^{n} P(T = t_i, D = \delta_i) = \prod_{i=1}^{n} p_i \ . \tag{1}$$

Remark: here p_i is the joint probability for (T_i, δ_i).

This somewhat awkward 2 dimensional distribution can equivalently be written as two 1-dimensional sub-distributions or sub-survival functions

$$U_1(t) = P(T > t, \delta = 1) \quad \text{and} \quad U_0(t) = P(T > t, \delta = 0). \tag{2}$$

By equivalent we mean 1-1 correspondence, i.e. given the 2-dimensional joint distribution, we can derive the two sub-survival functions. Conversely, given the two sub-survival functions, we can determine the joint distribution.

Using these two sub-survival functions the likelihood can be written as

$$\prod_{i=1}^{n} \{-\Delta U_1(T_i)\}^{\delta_i} \{-\Delta U_0(T_i)\}^{1-\delta_i}.$$

Here $\Delta U_1(t) = U_1(t+) - U_1(t-) = -P(T = t, \delta = 1)$ etc.

Until now we only assumed iid-ness of the (T_i, δ_i) vectors, or patient to patient iid.

If we further assume the random independent censoring model, that is, the ‘plus’ sign is due to an independent follow-up time C_i been shorter then the associated lifetime X_i:

$$T_i = \min(X_i, C_i) \quad \text{and} \quad \delta_i = I[X_i \le C_i]$$

then we have

$$U_1(t) = P(T > t, \delta = 1) = \int_t^\infty 1 - G(s) dF(s) \tag{3}$$

and

$$U_0(t) = P(T > t, \delta = 0) = \int_t^\infty 1 - F(s) dG(s) \ , \tag{4}$$

in the above we used $G(\cdot)$ to denote the CDF of C_i and $F(\cdot)$ to denote the CDF of X_i.

It can be shown that the two sub-survival functions, U_0 and U_1, are also equivalent to the two proper distributions, $F(s)$ and $G(s)$ under the independent random censorship model.

Exercise: Given the two sub-survival function U_1 and U_0, derive the distributions F and G under independent random censorship model. [The converse is already given as in the above two integrals: determine the two sub-survival functions, U_1 and U_0, in terms of F and G.]

[Remark: identifiability. If we do not assume independency then there are many choices of F and G and with various degrees of dependency and all of them will give the same sub-survival functions U_1 and U_0. See Tsiatis]

Assume independent censoring and given n iid observations, (T_i, δ_i), the likelihood function can further be written in terms of F and G:

$$\prod_{i=1}^n [(1 - G(t_i)) dF(t_i)]^{\delta_i} [(1 - F(t_i)) dG(t_i)]^{1-\delta_i} \ . \tag{5}$$

In most cases, we are mainly concerned with the inference of $F(\cdot)$; not $G(\cdot)$, since the G contains information related to the quality of follow-up time. Only F contains the information about survival. So, the terms that do not involve F are considered constant, and the likelihood (concerning F) is proportional to

$$\prod_{i=1}^n [dF(t_i)]^{\delta_i} [1 - F(t_i)]^{1-\delta_i} \ . \tag{6}$$

The log likelihood is, up to a constant term,

$$\sum_i \delta_i \log dF(t_i) + (1 - \delta_i) \log[1 - F(t_i)] \ . \tag{7}$$

If all the observations are completely observed, i.e. no censor (all $\delta = 1$) then the likelihood function is just

$$\prod_i [dF(t_i)] \ . \tag{8}$$

This last likelihood often written as

$$\prod p_i \quad \text{with} \quad \sum p_i \le 1.$$

Remark: Notice the different meanings of the p_i here and in (1) where the p_i is for the 2-dimensional joint probabilities.

When there are some censored observations, the maximization of (7) is obtained by a discrete distribution, the so called Kaplan-Meier estimator (Kaplan-Meier 1958).

Remark: If the observed data is merely independent but not identically distributed (like in a regression setting), then the final (log) likelihood is similar to (6) or (7) except the CDF $F(\cdot)$ will become $F_i(\cdot)$.

0.1 Likelihood in Terms of Hazard

Because cumulative hazard function and cumulative distribution function has 1 to 1 correspondence, this censored empirical likelihood can be written in terms of the hazard. However, due to the continuous/discrete version of the formula, we may end up with several versions of the likelihood.

Poisson version of the likelihood:

Let $H(t)$ denote the cumulative hazard function for the random variable X_i. By using the relation

$$1 - F(t) = e^{-H(t)}, \tag{9}$$

for a single observation (T_i, δ_i), the contribution to log empirical likelihood is

$$\delta_i \log dF(t_i) + (1-\delta_i)\log[1-F(t_i)] = \delta_i \log \Delta H(t_i) - H(t_i) \ . \tag{10}$$

However, the CDF F that achieve the maximum of log likelihood, the Kaplan-Meier estimator, is discrete. If we assume purely discrete H then the function value at t is just the sum of all jumps up to t:

$$\delta_i \log \Delta H(t_i) - \sum_j \Delta H(t_j) I[t_j \le t_i]. \tag{11}$$

For the log likelihood based on n iid observations, the log likelihood is the sum of n terms each like the above. And if merely independent is assumed, the likelihood will be H_i

Remark: careful reader may detect some inconsistency: we used a formula (9) which only works for the continuous case when deriving (10) but later assumes a purely discrete H in going from (10) to (11). If the underlying distribution for the survival times are continuous, then the jumps in the Kaplan-Meier etc. are shrinking to zero as sample size grow, and the inconsistency should diminish. But if the true survival distribution is discrete, then the discrepancy will stay even as sample size grow.

Bionomial version of the likelihood:

Here we always stick to the discrete version of the CDF/hazard function.

By using the relation

$$1 - F(t) = \prod_{s \le t} [1 - \Delta H(s)]$$

for a single observation observation T_i, δ_i, the contribution to the log empirical likelihood is,

$$\delta_i \log \Delta H(t_i) + \sum_j I[t_j < t_i] \log[1 - \Delta H(t_j)] + (1-\delta_i) \log[1 - \Delta H(t_i)]$$

Notice the last term is always zero, (assuming that the cumulative hazard function $H(\cdot)$ do not jump when $\delta = 0$). Thus we have

$$\delta_i \log \Delta H(t_i) + \sum_j I[t_j < t_i] \log[1 - \Delta H(t_j)].$$

The likelihood for n iid observations is then

$$\sum_{i=1}^n \delta_i \log \Delta H(t_i) + \sum_i \sum_j I[t_j < t_i] \log[1 - \Delta H(t_j)].$$

Switch the order of summation on the second term, we have

$$\sum_{i=1}^n \delta_i \log \Delta H(t_i) + \sum_j \log[1 - \Delta H(t_j)] \left(\sum_i I[t_j < t_i] \right)$$

Finally, use different subscript for the double summation: use t_s for t_i and use t_i for t_j, the log empirical likelihood for n observations is

$$\sum_{i=1}^n \delta_i \log \Delta H(t_i) + \sum_{i=1}^n \log[1 - \Delta H(t_i)] \left(\sum_s I[t_i < t_s] \right) \quad . \tag{12}$$

One interesting observation is that the hazard empirical likelihood using either the Poisson or binomial version, the maximizer is always the Nelson-Aalen estimator:

$$\Delta H^*(t_i) = \frac{\sum_j \delta_i I[t_j = t_i]}{\sum_j I[t_i \le t_j]}.$$

This can be verified by taking derivative and set it to zero in the log likelihood expression.

However, the Poisson version is easier to work with when we impose Cox model type of constraints.

When we use parametric approach to the analysis of censored data, the CDF/hazard function are usually continuous and thus we use likelihood (7) or (10). When we use nonparametric approach the maximizer are usually discrete and thus we use the likelihood (7), (11) or (12), depending if we are modeling the hazard or mean.