Left-Truncated Right-Censored Data Maximum Likelihood & Nelson-Aalen Estimators

As we discussed in class, left-truncated and right-censored survival data should be viewed as the set of observations $\{(T_i,\, \Delta_i,\, L_i) :\ 1 \le i \le M,\ T_i \le L_i\}$, where M is a large (nonrandom) index-number of 'potential observations' reflecting an underlying population, but we can only observe $(T_i,\, \Delta_i,\, L_i)$ for $L_i \le T_i$. Here we should think of underlying *latent* independent variables $(L_i,\, X_i,\, C_i,$ where the statistical parameters of interest are only those associated with the density $f_X(\cdot, \vartheta)$ of the failure time X_i, not those of the distribution for the left-truncation time L_i or the right-censoring time C_i.

The Likelihood for the observed data (after striking out the factors that do not depend upon ϑ), is

$$\mathrm{Lik}(\vartheta) \ = \ \prod_{i: T_i \ge L_i} \Big[\Big(\frac{f_X(T_i, \vartheta)}{S_C(T_i, \vartheta)} \Big)^{\Delta_i} \ \Big(\frac{S_X(T_i, \vartheta)}{S_X(L_i, \vartheta)} \Big)^{1-\Delta_i} \Big]$$

(*But note that in class, I inadvertently left out the denominator in the second factor in this product.*) Using the formulas we developed in class for hazard intensity and cumulative hazard, we can re-express the last likelihood formula as

$$\mathrm{Lik}(\vartheta) \ = \ \prod_{i: T_i \ge L_i} \Big[\, \lambda_X(T_i, \vartheta)^{\Delta_i} \, \exp\Big(- \int_{L_i}^{T_i} \lambda_X(s, \vartheta)\, ds \Big) \Big]$$

Therefore,

$$\begin{aligned} \mathrm{logLik}(\vartheta) \ &= \ \sum_{i: T_i \ge L_i} \Big[\Delta_i \, \log \lambda_X(T_i, \vartheta) \ - \ \int \, I_{[L_i < s \le T_i]} \, \lambda_X(s, \vartheta)\, ds \Big] \\ &= \ \int \, \Big\{ \log \lambda_X(s, \vartheta)\, dN(s) (- Y(s)\, \lambda_X(s, \vartheta)\, ds \Big\} \end{aligned}$$

where (in the left-truncated right-censored setting) N and Y are re-defined as

$$\left\{ \begin{array}{rcl} N(t) & = & \sum_i \, I_{[T_i \ge L_i]} \, I_{[T_i \le t]} \, \Delta_i \\ Y(t) & = & \sum_i \, I_{[T_i \ge t > L_i]} \end{array} \right.$$

Thus we find in all left-truncated right-censored survival datasets, that the MLE $\hat{\vartheta}$ solves the equation

$$\int \nabla_\vartheta \log \lambda_X(s,\vartheta) \Big\{ dN(s) \,-\, Y(s)\, \lambda_X(s,\vartheta)\, ds \Big\} \;=\; 0$$

When the hazard intensity is parameterized piecewise-constant by $\vartheta = (\lambda_1, \lambda_2, \ldots, \lambda_K)$ so that, for fixed and known $0 = a_0 < a_1 < \cdots < a_K$,

$$\lambda_X(s,\vartheta) \;\equiv\; \lambda_j \qquad \text{for all} \quad s \in (a_{j-1}, a_j]$$

we express the MLE's as

$$\hat{\lambda}_j \;=\; \int_{a_{j-1}}^{a_j} dN(s) \Big/ \int_{a_{j-1}}^{a_j} Y(s)\, ds$$

so that

$$\hat{\Lambda}_X(t) \;=\; \sum_{j:\, a_{j-1} \le t} \frac{a_j \wedge t \,-\, a_{j-1}}{\int_{a_{j-1}}^{a_j} Y(s)\, ds} \int_{a_{j-1}}^{a_j} dN(s)$$

Now assume that there are no tied survival times (as would be the case with probability 1 if the failure r.v. has a density). We complete the picture of estimating a flexibly parameterized cumulative hazard by observing that as the partition of the positive time-line by $\{a_j\}_{j=1}^K$ is made finer and finer (extending at least as far to the right as the largest survival time, all increments $N(a_j) - N(a_{j-1})$ become either 0 or 1, and the formula just given for $\hat{\Lambda}_X(t)$ has the limit

$$\hat{\Lambda}_X(t) \;=\; \int_0^t \frac{I_{[\max_i T_i \ge s]}}{Y(s)}\, dN(s)$$

This estimator is called the Neslon-Aalen nonparametric cumulative hazard estimator, and we will discuss its properties in class.