STAT 460/560 Class 17: Time-to-event data and survival analysis

Ben Bloem-Reddy

Reading: Chapter 9, [EH21].

Up to now in this course, we have assumed that we observe everything necessary to perform inference. Today we'll encounter a new type of data, which often has some sort of missingness.

1. Time-to-event data

Suppose we have data $T_1, \dots, T_n$ taking values in $\mathbb{R}_+$, and representing the time between some "starting event" (say, birth or the start of an experimental trial) and a pre-defined event of interest (death or the end of the trial). Sometimes we're lucky and all the event times are observed. Suppose for now that that is the case.

Suppose the (unknown) CDF of the T_i's is F, with PMF/PDF f. The **survival function** is

$$S(t) = P(T > t) = \begin{cases} \int_t^\infty f(s)\, ds & T \text{ continuous} \\ \sum_{t_k \geq t} f_{t_k} & T \text{ discrete.} \end{cases} \tag{17.1}$$

The **hazard rate** can be interpreted as the instantaneous rate of event occurrence, conditioned on the event occurring at least at time t,

$$h(t) = f(t)/P(T \geq t) \; . \tag{17.2}$$

The hazard rate can be used to obtain the survival function. In the discrete case, for $t \in [t_j, t_{j+1})$ this is

$$\begin{aligned} S(t) &= P(T > t) = P(T > t_j) = P(T > t_j | T > t_{j-1}) P(T > t_{j-1}) \\ &= P(T > 0) \prod_{k : t_k \leq t} P(T > t_k | T > t_{k-1}) \\ &= \prod_{k : t_k \leq t} P(T > t_k | T \geq t_k) \\ &= \prod_{k : t_k \leq t} (1 - h(t_k)) \end{aligned}$$

Also observe that in the discrete case,

$$\frac{S(t_j)}{S(t_{j-1})} = 1 - h(t_j) \; . \tag{17.3}$$

In the continuous case, the instantaneous rate of event occurrence at time t is

$$h(t)dt := P(T \in (t, t + dt) \mid T > t) \; .$$

To obtain an expression for $S(t)$, consider partitioning the interval $[0, t]$ into N sub-intervals of length dt_N, each centered at t_k. Using this discretization, assuming that $h(t_k)dt_N < 1$ (which it will be eventually as

$dt_N \to 0$ as $N \to \infty$), we can write

$$\begin{aligned} S_N(t) &= \prod_{k: t_k \le t} (1 - h(t_k) dt_N) \\ &= \exp\left(\sum_{k: t_k \le t} \ln(1 - h(t_k) dt_N) \right) \\ &= \exp\left(- \sum_{k: t_k \le t} h(t_k) dt_N + \frac{1}{2} h(t_k)^2 (dt_N)^2 + O(dt_N^3) \right) \\ &\to \exp\left(- \int_0^t h(s) ds \right) = P(T > t) = S(t) \ , \end{aligned}$$

as $N \to \infty$. In analogy to the discrete case, for $u < t$,

$$\frac{S(t)}{S(u)} = \exp\left(- \int_u^t h(s) ds \right) \ .$$

2. Censored data and Kaplan–Meier curves

Often, not all event times are observed—the trial may end, a subject may drop out of the trial, and so on. These are examples of **right censoring** (sometimes just referred to as censoring): $T_j > t_j^*$, where t_j^* is the time at which subject j is no longer observed, but T_j has not yet been observed. In such cases, data are recorded as $Z_i = (T_i, C_i)$, where

$$C_i = \begin{cases} 0 & \text{if event of interest is observed; and} \\ 1 & \text{if censored.} \end{cases} \tag{17.4}$$

Consider the randomized clinical trial from the Northern California Oncology Group (NCOG) clinical trial, as in Ch. 9.2 of [EH21]. Patients in Arm A of the trial were treated with chemotherapy; patients in Arm B received chemotherapy and radiation. There is right-censoring in this data. If there weren't, we could estimate the survival function of each treatment arm as a (linear) functional of the empirical CDF:

$$\hat{S}_A(t) = 1 - \hat{F}_A(t) \ . \tag{17.5}$$

However, this is a biased estimate. Why? Let's write down the likelihood of our observations. Assuming that the event times are IID, the likelihood is

$$\mathcal{L}_n(z_1, \ldots, z_n; F) = \prod_{i=1}^n f(t_i)^{1-c_i} S(t_i)^{c_i} \ . \tag{17.6}$$

To work with this for estimation, we can simplify a few things. First, sort the event times in increasing order, so that $t_{(1)} \le t_{(2)} \le \cdots \le t_{(n)}$. Second, denote by $\mathcal{R}_j$ the **risk set** at $t_{(j)}$, i.e., those patients who have not exited the study up to and including time $t_{(j)}$. That is, $\mathcal{R}_j = \{i \colon t_i \ge t_{(j)}\}$. Let d_j denote the number of events at time $t_{(j)}$, so that $d_j / |\mathcal{R}_j|$ is the proportion of the risk set that exits via the event of interest at time $t_{(j)}$.

Activity 17.1. Show that the likelihood can be written

$$\mathcal{L}_n(z_1, \ldots, z_n; F) = \prod_{j: \text{distinct } t_{(j)}} h(t_{(j)})^{d_j} (1 - h(t_{(j)}))^{|\mathcal{R}_j| - d_j} \ . \tag{17.7}$$

Solution: The likelihood in (17.6) can be rewritten as

$$\mathcal{L}_n(z_1,\ldots,z_n;F) = \prod_{j=1}^{n} h(t_{(j)})^{1-c_{(j)}} S(t_{(j-1)})^{1-c_{(j)}} S(t_{(j)})^{c_{(j)}} \ ,$$

where $S(t_{(j-1)}) = P(T \geq t_{(j)})$ comes from the definition $h(t_{(j)}) = f(t_{(j)})/P(T \geq t_{(j)})$.

Using the identity (17.3), this becomes

$$\begin{aligned}\mathcal{L}_n(z_1,\ldots,z_n;F) &= \prod_{j=1}^{n} h(t_{(j)})^{1-c_{(j)}} S(t_{(j-1)})(1-h(t_{(j)}))^{c_{(j)}} \\ &= \prod_{j=1}^{n} h(t_{(j)})^{1-c_{(j)}} (1-h(t_{(j)}))^{c_{(j)}} \prod_{i=1}^{j-1}(1-h(t_{(i)})) \\ &= \prod_{j:\text{distinct } t_{(j)}} h(t_{(j)})^{d_j}(1-h(t_{(j)}))^{|\mathcal{R}_j|-d_j} \ .\end{aligned}$$

Hence, if we condition on the observed event times (whether censored or uncensored) and require our estimated CDF to only put non-zero probability mass to the uncensored times, the likelihood is

$$\mathcal{L}_n(z_1,\ldots,z_n;F) = \prod_{j:\text{distinct } t_{(j)}} h(t_{(j)})^{d_j}(1-h(t_{(j)}))^{|\mathcal{R}_j|-d_j} \ , \tag{17.8}$$

and the MLE for the hazard rate is found to be

$$\hat{h}(t) = \frac{d_j}{|\mathcal{R}_j|} \ , \quad t_{(j)} \leq t < t_{(j+1)} \ . \tag{17.9}$$

Plugging this into the survival function,

$$\hat{S}(t) = \prod_{i=1}^{j}(1-\hat{h}(t_{(i)})) = \prod_{i=1}^{j}\left(1-\frac{d_i}{|\mathcal{R}_i|}\right) \ , \quad t_{(j)} \leq t < t_{(j+1)} \ . \tag{17.10}$$

This is called the **Kaplan–Meier** (KM) estimator of the survival function.

For comparison, the KM estimator and the estimator obtained from the empirical CDF ignoring censoring as in (17.5) are shown in Fig. 1.

Activity 17.2. Compare the KM estimator (17.10) with the estimator one would get by ignoring censoring and just using the empirical CDF (as in (17.5)). Show that if there is no censoring then the two estimators are equal.

Hint: Let $\bar{d}_j = \sum_{i=1}^{j} d_i$ (with $\bar{d}_0 = 0$), and note that if there is no censoring then $|\mathcal{R}_j| = n - \bar{d}_{j-1}$.

Solution: Without censoring, the KM estimator becomes, for $t_{(j)} \leq t < t_{(j+1)}$,

$$\begin{aligned}\hat{S}_n(t) = \prod_{i=1}^{j}\left(1-\frac{d_i}{|\mathcal{R}_i|}\right) &= \prod_{i=1}^{j}\left(\frac{n-\bar{d}_{i-1}-d_i}{n-\bar{d}_{i-1}}\right) \\ &= \prod_{i=1}^{j}\left(\frac{n-\bar{d}_i}{n-\bar{d}_{i-1}}\right) \\ &= \frac{n-\bar{d}_j}{n} = 1-\bar{d}_j/n = 1-\hat{F}_n(t)\end{aligned}$$

This holds more generally over any interval (u,v) on which there is no censoring. With this in mind, one way to interpret the KM estimator is that it is the empirical CDF estimator of the survival function, except that the underlying empirical population changes every time there is a censoring event.

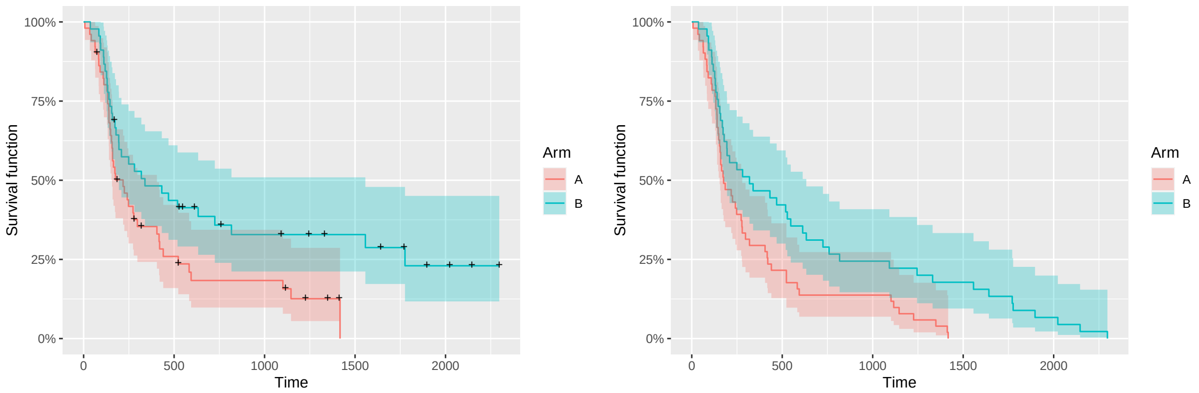

Figure 1: Estimated survival functions. Left: KM estimator. Right: Empirical CDF-based estimator ignoring censoring.

3. Proportional hazards

The KM estimator is non-parametric; it assumes nothing about the underlying probability model beyond basic IID assumptions and that F has discrete support (and the latter was just for convenience). Perhaps we are willing to make stronger assumptions, or perhaps we have covariates that we would like to use in our model. If the covariates aren't discrete and relatively few, the covariate-wise KM estimators aren't very appealing.

The most widely used *non*-non-parametric model for time-to-event data is the (Cox) **proportional hazards** model. It assumes that the i-th individual has the hazard rate

$$h_i(t) = h_0(t)e^{x_i\beta} \; . \tag{17.11}$$

The "proportional" in the name comes from the fact that if we look at the hazard ratio of two individuals, we see that it is constant (i.e., "proportional")

$$\frac{h_i(t)}{h_j(t)} = e^{(x_i - x_j)\beta} \; . \tag{17.12}$$

If we condition on the risk set $\mathcal{R}_j$ at each $t_{(j)}$, then the **partial likelihood**[1] is (assuming no ties)

$$\mathcal{L}_n(z_1, \ldots, z_n; \beta) = \prod_{j=1}^{n} \frac{e^{x_{(j)}\beta}}{\sum_{i \in \mathcal{R}_j} e^{x_i\beta}} \; . \tag{17.13}$$

This can be optimized numerically (Newton–Raphson/Fisher scoring), and the usual asymptotic normality arguments apply (with a caveat; see below) for finding approximate confidence intervals, etc.

Note that the baseline hazard function $h_0(t)$ doesn't play a role in estimation. That's because it's shared by all of the individuals, and it cancels out of the ratios in the likelihood function. Technically, that makes this model semi-parametric: the parametric log-linear model for the individual hazard rates multiplied by the arbitrarily complex (non-parametric) baseline hazard. In most instantiations of semi-parametric models, both components need to be estimated (with the non-parametric nuisance the more challenging one to estimate well); here, the math works out nicely so that we can essentially ignore the non-parametric part. If we did have to estimate both parts then the asymptotics/approximate confidence intervals/statistical theory get very complicated. See [van98], Ch. 25 for more.

[1] Partial here because we're not modeling the censoring process.

References

[EH21] B. Efron and T. Hastie. *Computer Age Statistical Inference*. Cambridge University Press, 2021.

[van98] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.