Emory Rollins School of Public Health | Department of Biostatistics and Bioinformatics

BIOS 522: Survival Analysis Methods

Reading 10:

Left truncation and interval censoring

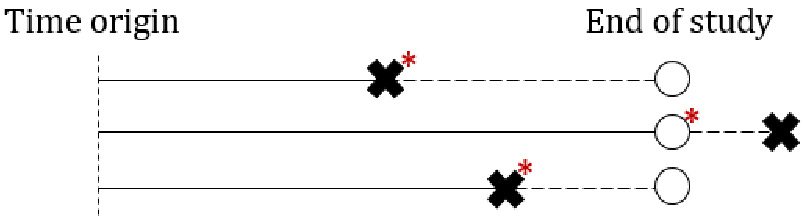
This week, we will discuss two key data structures for survival data extending beyond traditional right censoring. We consider left truncation and interval censoring.

Part 1. Delayed/staggered entry and left truncation

Right-censoring review

Censoring concerns the observation of failure times. The typical way this arises is due to the individual not having failed by the time the data are analyzed.

Consider the example below with three individuals. Failure times are marked with black Xs, and censoring times are marked with open circles. We observe the failure times of the first and last individuals because these occur before the end of the study. We do not observe the failure time of the middle individual because it occurs after the end of the study. We say this middle observation is **right-censored** because we do not observe anything that occurs to the right of the censoring time.



Importantly, we do know that this person survived at least until the end of the study. Thus, this person contributes information to the Kaplan-Meier estimate of survival, log-rank test, Cox proportional hazards regression model, etc.

Delayed entry

We have assumed throughout this course that all individuals enter the study at the relevant time origin. If the time origin is time of surgery, all patients are observed from the time of surgery onwards. Thus, we would capture any failures that occurred shortly after surgery. In practice, because we can define our time origin in many different ways, individuals may enter a study after the time origin. This is known as **delayed entry**.

Consider a study of pregnancy outcomes. The primary endpoint is miscarriage. Pregnant women receiving prenatal care at study clinics are invited to enroll. Imagine we set the time origin as last menstrual date, but women do not access prenatal care until 7 weeks gestation. Note that we will not observe miscarriages that occur between conception and 7 weeks gestation, particularly as women may have early miscarriages without knowing that they were ever pregnant.

Imagine now that we use our study data to estimate survival $S(t)$ as a function of gestational age t. Since we do not observe any miscarriages before 7 weeks gestation, 100% of study participants will have pregnancies that last through 7 weeks gestation. The estimated probability of miscarriage before 7 weeks will be 0%. Of course, we realize this is biased as early miscarriages can occur. We have simply failed to observe them in our data set. (This is closely related to the epidemiological concept of **immortal person-time bias**.)

A better approach would be to set the time origin to 7 weeks gestation. Then, we will estimate survival $S(t)$ over time only among pregnancies that were viable at 7 weeks gestation. Survival starts at 100% at 7 weeks and declines over time.

Staggered entry

In the example above, all women enrolled in the study at 7 weeks gestation. In practice, women will access prenatal care at different times. Some women will not know they are pregnant until later (e.g. 10 or 12 weeks gestation). Even if we move the study origin from conception to 7 weeks, many women will still enter the study at times after the study origin. This is known as **staggered entry**.

In the same way that delayed entry can lead to bias, staggered entry can also lead to bias. We will tend to miss early miscarriages, particularly in women with late entry times.

Example – Long-term outcomes of atomic bomb survivors

A notable historical example is a study of long-term outcomes in survivors of the atomic bomb dropped in Hiroshima in 1945. Starting in 1968, the local government urged survivors to register for the study. Thus, study participants had to have survived until at least 1968 to participate in the study, so participants had longer survival times on average than the underlying population. The earliest possible study entry time was 23 years after the bombing.

Furthermore, not all eligible individuals enrolled in 1968. New participants were enrolled over time through the end of the study in 1992. Thus, there is potential for bias because shorter survival times will be under-represented.

Notation for staggered entry

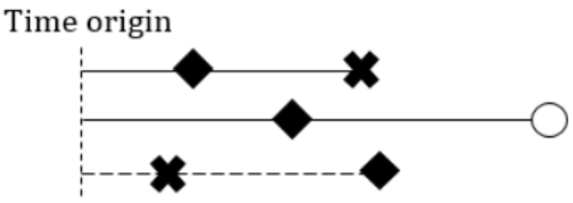
In addition to having a survival (or censoring time) T_i^* and an event indicator δ_i, each individual has an **entry time** S_i. The entry time is the time, relative to the origin, when the individual enters the study. $S_i = 0$ for individuals who enter at the time origin, otherwise $S_i > 0$.

In our pregnancy study example, if the time origin is 7 weeks gestation, and a woman enters the study at 9 weeks gestation, her entry time is $S_i = 2$. If the woman has a miscarriage at 11 weeks gestation, her survival time is $T_i^* = 4$ (4 weeks after our new time origin) and $\delta_i = 1$.

Left truncation

Importantly, we are only able to observe individuals whose entry time is *before* their failure time. Otherwise, these individuals would fail before they are ever enrolled, so they are missed entirely. Individuals who are not included in the study because their failure time occurred before their entry time are referred to as **left-truncated**. Their failure time is to the *left* of their entry time.

In the figure below, hypothetical data is plotted for three individuals. Black Xs are failure times T_i, white circles are censoring times C_i, and black diamonds are study entry times S_i. All individuals have study entry times after the time origin. The third individual is left-truncated because their failure time is to the left of their entry time. They would never enter the study.



Continuing with our pregnancy study example, if the time origin is 7 weeks gestation, and a woman would enter the study at 10 weeks gestation, but she has a miscarriage at 8 weeks gestation, she would not be included in the study.

There is an important distinction between truncated and censored data. When data are right-censored, we know that an individual has been right censored, though we do not know the survival time (just that it is larger than the censoring time!). They contribute to the risk set/denominator in our calculations. In contrast, when data are left-truncated, we do not even know that an individual *exists*. Thus, we cannot include them in our calculations, and this may limit what we can estimate. In the Hiroshima example, we cannot estimate how survival declined between 0 and 23 years after the bombing using the long-term survival study. We can only estimate survival *beyond* 23 years, conditional on having survived at least 23 years.

Left truncation and Kaplan-Meier estimation

We can accommodate staggered entry – where people enter into the study at different times after the origin -- into our data analysis methods, as long as some individuals are observed during all time periods. Recall that for the Kaplan-Meier and log-rank tests, we break time into intervals. When we have staggered entry, individuals contribute to the risk set/denominator of intervals only *after* they enter the study.

Example: Consider a small hypothetical data set of patients with Ebola virus disease. The time origin is the time when symptoms first appear. Some cases are detected several days after their symptoms appear. Half of the participants enter the study on their date of symptom onset ($S_i = 0$), and half enter later. Some early deaths may have been missed as a result. The outcome is death. Individuals are followed for 28 days before they are censored.

Entry time S_i	**Follow-up time T_i^***	**Event indicator δ_i**
0	2	1
2	4	1
4	6	1
0	6	1
2	8	1
4	8	1
0	28	0
2	28	0
4	28	0
0	28	0
2	28	0
4	28	0

The Kaplan-Meier estimate of $S(t)$ is calculated below. The number at risk n_j at time t_j is calculated as the total number of participants i who have previously entered the study ($S_i < t_j$) and have not yet failed or been censored ($T_i^* \geq t_j$). Note that the number at risk in each interval n_j increases until the last study entry time as new participants are enrolled at times 2 and 4 days.

Unique failure/ censoring time t_j	Number at risk n_j during $(t_{j-1}, t_j]$	Number of deaths d_j at t_j	Conditional survival probability $\hat{q}_j$	Kaplan-Meier estimate in $[t_j, t_{j+1})$
$t_0 = 0$				$t = [0,2)$ $\hat{S}(t) = 1$
$t_1 = 2$	$t = (0,2]$ $n_1 = 4$	$d_1 = 1$	$\hat{q}_1 = \left(1 - \frac{1}{4}\right)$	$t = [2,4)$ $\hat{S}(t) = 0.750$
$t_2 = 4$	$t = (2,4]$ $n_2 = 7$	$d_2 = 1$	$\hat{q}_2 = \left(1 - \frac{1}{7}\right)$	$t = [4,6)$ $\hat{S}(t) = 0.643$
$t_3 = 6$	$t = (4,6]$ $n_3 = 10$	$d_3 = 2$	$\hat{q}_3 = \left(1 - \frac{2}{10}\right)$	$t = [6,8)$ $\hat{S}(t) = 0.514$
$t_4 = 8$	$t = (6,8]$ $n_4 = 8$	$d_4 = 2$	$\hat{q}_4 = \left(1 - \frac{2}{8}\right)$	$t = [8,28)$ $\hat{S}(t) = 0.386$
$t_5 = 28$	$t = (8,28]$ $n_5 = 6$	$d_5 = 0$	$\hat{q}_5 = 1$	$t = [28, \infty)$ $\hat{S}(t) = 0.386$

Left truncation and the partial likelihood

Similarly, we can accommodate left truncation in our partial likelihood calculations. Recall the form of the partial likelihood for a Cox model with one covariate:

$$L(\beta) = \prod_{i=1}^{m} \frac{\exp(\beta X_i)}{\sum_{j \in R(T_i)} \exp(\beta X_j)}$$

Like for Kaplan-Meier estimation, the risk set $R(T_i)$ at unique failure time T_i only includes individuals j who have previously entered the study, ($S_j < T_i$) and have not yet failed or been censored ($T_j^* \geq T_i$).

A note on left censoring

Though relatively uncommon, data can also be **left-censored**. Unlike truncated data, when data are left-censored we know of the existence of observations that occurred before some time, but do not know the exact time when the event occurred.

Example: Consider a study focusing on time to recurrence of a cancer after surgical removal of the primary tumor. Three months after surgery, patients are examined to see whether the cancer has recurred. For patients whose cancer had recurred by the time of the three-month appointment, the actual time to recurrence is less than three months. For these subjects, their recurrence time is left-censored.

Part 2. Interval censoring

Motivating examples

Another common type of data structure occurs when we do not know exactly when an event has occurred. Rather, we know that it occurred during some interval. For example, we might study the time until development of an asymptomatic kidney stone. As the kidney stone can only be detected by x-ray, a positive x-ray indicates that the kidney stone developed between the last x-ray and the current x-ray, but we do not know exactly when. We refer to this type of data as **interval-censored**.

Interval censoring is common in studies of non-lethal events that require specialized tests to determine that they have occurred. Other examples include:

- Time to onset of dementia as determined by cognitive testing
- Time to undetectable HIV viral load in AIDS studies
- Time to clearance of an infection measured by viral assay
- Time to ulcer as determined by endoscopy

Interval censoring notation

Imagine that we know that our survival time T_i lies within an interval (L_i, R_i) referring to the left-hand and right-hand bounds. Keeping with our asymptomatic kidney stone example, L_i is the time of *last negative* x-ray, and R_i is the time of the *first positive* x-ray. Thus, T_i is somewhere between these bounds.

Naïve approaches for analyzing interval censored data include:

- Setting T_i as the interval midpoint, $T_i = (L_i + R_i)/2$
- Setting T_i as a random point in the interval (L_i, R_i)
- If the event is defined by a continuous scores crossing a threshold, you can use the score values at L_i and R_i to interpolate when the threshold would have been crossed assuming they are connected by a straight line

These naïve approaches are biased, though sometimes they are used in practice. This bias is most apparent when the intervals are long relative to the study length and/or the intervals vary in length across time or across individuals.

Maximum likelihood for interval censoring

A more rigorous analysis uses a **maximum likelihood approach**. Briefly, suppose we have independent observations from n subjects, and that each subject has an left-hand bound L_i and right-hand bound R_i.

We know that the survival time T_i is in the interval $(L_i, R_i]$. If the individual is censored, then L_i is the last observation, and $R_i = \infty$.

The likelihood function for parameter(s) θ for n individuals is:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \Pr(L_i < T_i \leq R_i | \theta) = \prod_{i=1}^{n} \big(F(R_i|\theta) - F(L_i|\theta)\big)$$

There are then two ways to estimate our survival function $S(t)$.

(1) We can calculate a nonparametric estimate of $S(t)$. When each individual has a different set of intervals, this can be quite difficult to do. **Turnbull's nonparametric estimator** uses an iterative isotonic regression algorithm. This is one commonly used approach to generate basic survival plots for interval-censored data.

(2) We can assume a parametric distribution for T (e.g. exponential, Weibull). We can set $F(\cdot)$ as the CDF of our preferred distribution. We then maximize the likelihood to estimate our unknown parameters (e.g. exponential rate; Weibull rate, shape) and use these to estimate $S(t)$.

Imagine we assume the data follow an exponential distribution:

$$\begin{aligned}\mathcal{L}(\lambda) &= \prod_{i=1}^{n} \Big(\big(1 - e^{-\lambda R_i}\big) - \big(1 - e^{-\lambda L_i}\big)\Big) \\ &= \prod_{i=1}^{n} \big(e^{-\lambda L_i} - e^{-\lambda R_i}\big)\end{aligned}$$

For censored observations, the likelihood contribution is $e^{-\lambda L_i} - 0 = e^{-\lambda L_i}$ since the survival approaches zero as $L_i \rightarrow \infty$. Like any likelihood function, we can maximize this function with respect to the parameter λ.

The approach follows a similar logic for a distribution with more than one parameter, like the Weibull distribution, or for a parametric regression model, like the Weibull PH or AFT model. A pro of parametric models is that they can nicely accommodate interval censored data.

Informative censoring

As with right censored data, the concepts of noninformative and informative censoring apply with interval censored data. In general, one way to ensure that censoring is noninformative is to select the screening times in advance. Note these intervals do not need to be equally spaced.

In contrast, informative interval censoring could arise if patients who began to experience kidney pain went to their doctor and, as a result, received a diagnostic test.

Part 3. Looking ahead

Next week we will turn our attentions to clinical trials and some of the unique features of survival analysis in trials. Of particular importance is calculating sample size requirements for trial planning.