Intro to Survival Analysis

STA 198: Introduction to Health Data Science

Yue Jiang

July 21, 2020

The following material was used by Yue Jiang during a live lecture.

Without the accompanying oral comments, the text is incomplete as a record of the presentation.

Goal

Survival analysis is a complex topic, and you are strongly encouraged to take a survival analysis course if you plan to analyze data of this type. The goal of our coverage is to give you the skills you need to understand results of simple descriptive statistics in this setting, and we will not have time to discuss more complex modeling of survival data in this course.

Survival data

In many studies, the outcome of interest is the amount of time from an initial observation until the occurrence of some event of interest, e.g.

- Time from transplant surgery until new organ failure
- Time to death in a pancreatic cancer trial
- Time to menopause
- Time to divorce
- Time to receipt of PhD

Typically, the event of interest is called a *failure* (even if it is a good thing). The time interval between a starting point and the failure is known as the *survival time* and is often represented by t.

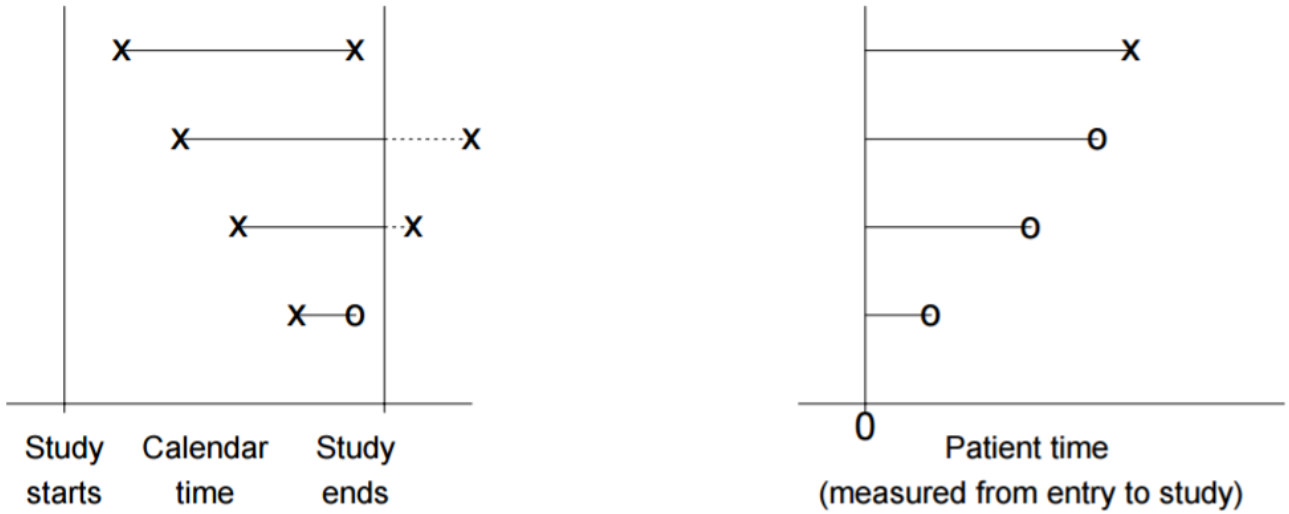
Survival data

Certain aspects of survival data make data analysis particularly challenging.

- Typically, not all the individuals are observed until their times of failure
  - An organ transplant recipient may die in an automobile accident before the new organ fails
  - A PhD student may withdraw from the program to start a multi-billion dollar health company
  - Not everyone gets divorced
  - A pancreatic cancer patient may move to Fiji instead of choosing to undergo further treatment
- In this case, an observation is said to be *censored* at the last point of contact with the patient.

Study time and patient time

It is important to distinguish between study time and patient time.



- A study may start enrolling patients in September and continue until all 500 patients have been enrolled
- This is likely to take months or years
- Time is typically converted to patient time (time between enrollment and failure or censoring) before analysis

Survival function

The distribution of survival times is characterized by the *survival function*, represented by $S(t)$. For a continuous random variable T,

$$S(t) = P(T > t),$$

and $S(t)$ represents the proportion of individuals who have not yet failed.

The graph of $S(t)$ versus t is called a survival curve. The survival curve shows the proportion of survivors at any given time. It is non-increasing, with $S(0) = 1$ and $\lim_{t \to \infty} S(t) = 0$.

Simple example

A small study enrolls 10 patients, whose outcomes are below:

Patient	Event Time (x)	Event Type
1	4.5	Death
2	7.5	Death
3	8.5	Censored
4	11.5	Death
5	13.5	Censored
6	15.5	Death
7	16.6	Death
8	17.5	Censored
9	19.5	Death
10	21.5	Censored

How do we estimate the survival curve for these data?

Kaplan-Meier estimate

Perhaps the most popular estimate of a survival curve is the *Kaplan-Meier* or *product-limit* estimate. This method is actually fairly intuitive.

First, define the following quantities.

- Y_t: # at risk of failure at time t (i.e., those who did not fail before t and those who were not censored before t)
- d_t: # who fail at time t
- $q_t = \frac{d_t}{Y_t}$: estimated probability of failing at time t
- $S(t)$: cumulative probability of surviving beyond time t, estimated as

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{Y_{t_i}}\right).$$

- The $\prod$ symbol is for multiplication: e.g., $\prod_{i=1}^{3} x_i = x_1 x_2 x_3$ and $\prod_{i=1}^{5} i = 1 \times 2 \times 3 \times 4 \times 5$.

How is that intuitive?

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{Y_{t_i}}\right).$$

At each time t, the probability of surviving is just $1 - P(\textit{failing})$. Before there are any failures in the data, our estimated $\hat{S}(t) = 1$. At the time of the first failure, this probability falls below 1 and is simply one minus the probability of failing at that time, or $1 - \frac{\#\textit{failures}}{\#\textit{at risk}}$.

After the first failure, things get more complicated. At the time of the second failure, you can calculate $1 - \frac{\#\textit{failures}}{\#\textit{at risk}}$, but this doesn't provide the whole picture, as someone else has already died. In fact, this is the conditional probability of surviving now that you've made it past the time of the first failure.

How is that intuitive?

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{Y_{t_i}}\right).$$

How do you then calculate the total (unconditional) probability of survival? That is just the product of the probability of surviving past the first failure times the conditional probability of surviving beyond the second failure given that you made it past the first:

$$\begin{aligned}
&P(\text{survived past first and second times}) \\
&= P(\text{survive past first time})P(\text{survive past second time}|\text{survived past first time}) \\
&= \left(1 - \frac{\#\text{failures, failure time 1}}{\#\text{at risk of failing, failure time 1}}\right)\left(1 - \frac{\#\text{failures, failure time 2}}{\#\text{at risk of failing, failure time 2}}\right)
\end{aligned}$$

If someone is censored, they are no longer at risk of failing at the next failure time and are taken out of the calculation.

Calculating the Kaplan-Meier estimate

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{Y_{t_i}}\right).$$

t	# Failed (d_t)	# Censored	# Left (Y_{t+1})	$\hat{S}(t)$
0.0	0	0		
4.5	1	0		
7.5	1	0		
8.5	0	1		
11.5	1	0		
13.5	0	1		
15.5	1	0		
16.5	1	0		
17.5	0	1		
19.5	1	0		
21.5	0	1		

Calculating the Kaplan-Meier estimate

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{Y_{t_i}}\right).$$

t	# Failed (d_t)	# Censored	# Left (Y_{t+1})	$\hat{S}(t)$
0.0	0	0	10	1
4.5	1	0		
7.5	1	0		
8.5	0	1		
11.5	1	0		
13.5	0	1		
15.5	1	0		
16.5	1	0		
17.5	0	1		
19.5	1	0		
21.5	0	1		

Calculating the Kaplan-Meier estimate

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{Y_{t_i}}\right).$$

t	# Failed (d_t)	# Censored	# Left (Y_{t+1})	$\hat{S}(t)$
0.0	0	0	10	1
4.5	1	0	9	$1 - \frac{1}{10} = 0.9$
7.5	1	0		
8.5	0	1		
11.5	1	0		
13.5	0	1		
15.5	1	0		
16.5	1	0		
17.5	0	1		
19.5	1	0		
21.5	0	1		

Calculating the Kaplan-Meier estimate

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{Y_{t_i}}\right).$$

t	# Failed (d_t)	# Censored	# Left (Y_{t+1})	$\hat{S}(t)$
0.0	0	0	10	1
4.5	1	0	9	0.9
7.5	1	0	8	$0.9 \times (1 - \frac{1}{9}) = 0.8$
8.5	0	1		
11.5	1	0		
13.5	0	1		
15.5	1	0		
16.5	1	0		
17.5	0	1		
19.5	1	0		
21.5	0	1		

Calculating the Kaplan-Meier estimate

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{Y_{t_i}}\right).$$

t	# Failed (d_t)	# Censored	# Left (Y_{t+1})	$\hat{S}(t)$
0.0	0	0	10	1
4.5	1	0	9	0.9
7.5	1	0	8	0.8
8.5	0	1	7	$0.8 \times (1 - \frac{0}{8}) = 0.8$
11.5	1	0		
13.5	0	1		
15.5	1	0		
16.5	1	0		
17.5	0	1		
19.5	1	0		
21.5	0	1		

Calculating the Kaplan-Meier estimate

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{Y_{t_i}}\right).$$

t	# Failed (d_t)	# Censored	# Left (Y_{t+1})	$\hat{S}(t)$
0.0	0	0	10	1
4.5	1	0	9	0.9
7.5	1	0	8	0.8
8.5	0	1	7	0.8
11.5	1	0	6	$0.8 \times (1 - \frac{1}{7}) = 0.69$
13.5	0	1		
15.5	1	0		
16.5	1	0		
17.5	0	1		
19.5	1	0		
21.5	0	1		

Calculating the Kaplan-Meier estimate

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{Y_{t_i}}\right).$$

t	# Failed (d_t)	# Censored	# Left (Y_{t+1})	$\hat{S}(t)$
0.0	0	0	10	1
4.5	1	0	9	0.9
7.5	1	0	8	0.8
8.5	0	1	7	0.8
11.5	1	0	6	0.69
13.5	0	1	5	0.69
15.5	1	0		
16.5	1	0		
17.5	0	1		
19.5	1	0		
21.5	0	1		

Calculating the Kaplan-Meier estimate

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{Y_{t_i}}\right).$$

t	# Failed (d_t)	# Censored	# Left (Y_{t+1})	$\hat{S}(t)$
0.0	0	0	10	1
4.5	1	0	9	0.9
7.5	1	0	8	0.8
8.5	0	1	7	0.8
11.5	1	0	6	0.69
13.5	0	1	5	0.69
15.5	1	0	4	$0.69 \times (1 - \frac{1}{5}) = 0.552$
16.5	1	0		
17.5	0	1		
19.5	1	0		
21.5	0	1		

Calculating the Kaplan-Meier estimate

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{Y_{t_i}}\right).$$

t	# Failed (d_t)	# Censored	# Left (Y_{t+1})	$\hat{S}(t)$
0.0	0	0	10	1
4.5	1	0	9	0.9
7.5	1	0	8	0.8
8.5	0	1	7	0.8
11.5	1	0	6	0.69
13.5	0	1	5	0.69
15.5	1	0	4	0.552
16.5	1	0	3	0.414
17.5	0	1	2	0.414
19.5	1	0	1	$0.414 \times (1 - \frac{1}{2}) = 0.207$
21.5	0	1		

Calculating the Kaplan-Meier estimate

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{Y_{t_i}}\right).$$

t	# Failed (d_t)	# Censored	# Left (Y_{t+1})	$\hat{S}(t)$
0.0	0	0	10	1
4.5	1	0	9	0.9
7.5	1	0	8	0.8
8.5	0	1	7	0.8
11.5	1	0	6	0.69
13.5	0	1	5	0.69
15.5	1	0	4	0.552
16.5	1	0	3	0.414
17.5	0	1	2	0.414
19.5	1	0	1	0.207
21.5	0	1	0	0.207

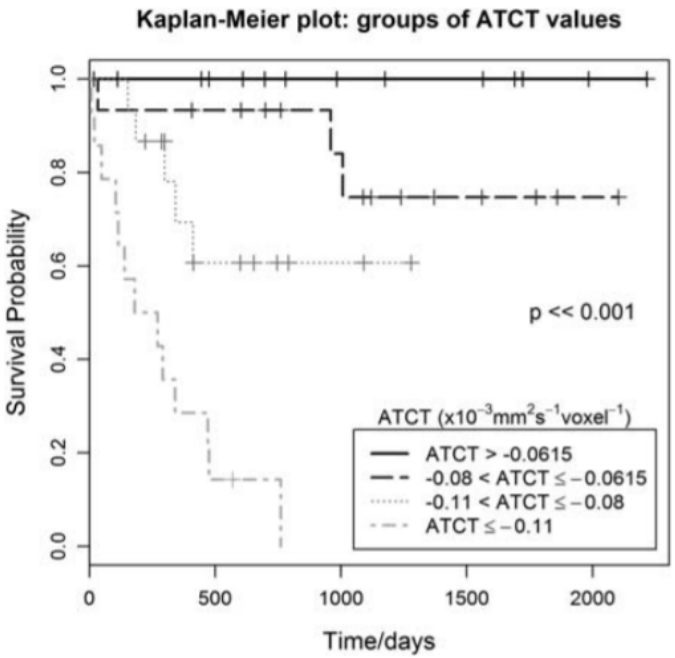
What would $\hat{S}(21.5)$ be if the last observation were a failure instead of censored?

KM estimate

In between failure times, the KM estimate does not change but is constant. This gives the estimated survival function its step-like appearance (we call this type of function a *step function*).

Tumors in children, 2012 *Neuro-oncology*

ATCT is an imaging-based biomarker of tumor prognosis



- Which biomarker values are associated with the best survival?
- Which values are associated with the worst survival?
- What is the median survival time in the group with the smallest ATCT values?
- If a child is in the group with the largest ATCT values, what is his/her estimated 5-year survival probability?

Log-rank test

How do we determine whether the difference in survival curves is statistically significant?

The log-rank test is quite intuitive. The idea behind it is to construct a 2×2 contingency table by group (assuming two groups; can be extended) at each time t at which a failure occurs. Then, these tables are combined in a specific way (Mantel-Haenszel) For this test, the null hypothesis is that the survival curves in the two groups are the same, e.g.:

$$H_0 : S_1(t) = S_2(t).$$

The test statistic follows a χ^2_1 distribution under the null hypothesis (if two groups).

Survival vs. hazard

- Survival function

$$S(t) = P(T > t)$$

- Hazard function

$$h(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

Instantaneous *failure rate* for observations, conditionally on already having survived to time t

- *Not* a probability (assuming continuous T)
- Non-negative and unbounded for all t
- One-to-one relationship between the survival and hazard functions (knowing one tells you exactly what the other is)

The Cox proportional hazard model

- By far the most commonly used regression model for survival data
- Flexible handling of covariates, with attractive interpretation
- Fairly easy to fit and widely implemented
- Assumes that there is a constant hazard ratio for two different subjects at all times. This may not always be satisfied!

Interpretation of Cox model estimates

$$h_z(t) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)$$

- Each β_j parameter is the log hazard ratio for a one-unit increase in the associated predictor, holding all other predictors constant
- Exponentiating, $\exp(\beta_j)$ is the hazard ratio between two individuals who values of X_j differ by one unit, holding all other predictors constant
- $h_0(t)$ is the baseline hazard, assumed common to all subjects

Note the similarity in spirit to interpretations from logistic regression. Positive $\hat{\beta}$ (or $\exp(\hat{\beta}) > 1$) implies greater relative hazard