Biostat 537: Survival Analaysis

TA Session 1

Ethan Ashby

September 23, 2023

Presentation Overview

Hi! I'm your TA!

My name is Ethan Ashby and I'm a third year PhD student in the Biostatistics department at UW.

Typically, office hours will contain a short lecture followed by time for questions.

Please feel free to email me (eashby (at) uw dot edu) with any questions as you progress through this course! Lecture slides will also be made available on my website.

Survival Data

Survival data is a special kind of outcome data denoting the *time until an event occurs*. Let T denote the positive random variable for a person's survival time.

Requires specifying a time origin ($t = 0$).

Some examples:

1. Hardware reliability study: time until component fails.
2. Vaccine study: time until infection with the flu.
3. Cancer study: time until disease progression.

Censoring

Censoring: when the survival time is not known precisely, but is known to lie before, after, or between certain values. Let $d = \{0, 1\}$ be a random variable indicating either censorship or an event respectively.

Right-censoring: when the true survival time is known to lie *after* a given value; i.e., $T \in [t_1, \infty)$.

Causes of right censoring

1. Study ends before participant experiences an event – aka administrative censoring.
2. Lost to follow up during study period.
3. Dropout from the study.

Censoring cont.

Left-censoring: when the true survival time is known to lie *before* a given time point; i.e., $T \in [0, t_1)$.

Interval-censoring: when the true survival time is known to lie *between* two time points; ; i.e., $T \in [t_1, t_2]$.

Censoring examples

An individual enrolls in a HIV surveillance study where they complete tests at regular monthly intervals. Assume the time origin ($t = 0$) denotes when the individual became sexually active and hence was at risk of acquiring HIV.

1. Individual tests positive for HIV at the first testing visit.
2. Individual reaches the end of the study without testing positive for HIV.
3. Individual tests negative at the first testing visit but positive at the second visit.

Left truncation

A form of *selection bias*, where we can only observe an event time if it is greater than a certain value.

Example: if we want to estimate time from cancer diagnosis to death and we recruit diagnosed patients, we may be excluding patients with small survival times who die before recruitment.

Survival data: Notation

For a given individual i, their survival data is summarized by the following vector $(t_i, d_i, X_{i1}, \ldots, X_{ip})$ where

1. $t_i = T_i \wedge C_i$ (minimum of survival and censoring time)
2. $d_i = \mathbb{I}(T_i < C_i)$ denotes whether the observation was observed or censored.
3. $(X_{i1}, \ldots, X_{ip})$ denote a vector of p-covariates.

Survival data: presentation

Survival data are often summarized like so: $(t_1, t_2, t_3, t_4+, t_5+)$ where "+" denotes censoring at the designated time.

To computers, we often encode survival data like so

Individual	t	d (failed or censored)
1	t_1	1
2	t_2	1
3	t_3	1
4	t_4	0
5	t_5	0

Survivor & Hazard Functions

Survivor function: $S(t) := P(T \geq t)$ gives the probability of surviving beyond a specified time t.

1. $S(t)$ is non-decreasing, $S(t = 0) = 1$ when $S(\infty) = 0$.

Hazard function: $h(t)$ describes the *instantaneous rate* (per unit time) of experiencing an event at time t *given* the individual survived up to time t. We formalize as follows:

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

Differences between survivor and hazard functions

	Survivor	Hazard
Who	Cohort	Individual
What	Probability	Event Rate
When	Cumulative (over time)	Instantaneous

Relationship between survivor & hazard functions

There exists a 1-1 relationship between survivor functions $S(t)$ and hazard functions $h(t)$.

$$S(t) = \exp\left[\underbrace{-\int_0^t h(u)du}_{\star}\right]$$

$$h(t) = -\left[\frac{dS(t)/dt}{S(t)}\right]$$

Note: $\star \equiv H(t)$ is often referred to as the *cumulative hazard function*.

Example: parametric survival model

Suppose the hazard function is constant over time with risk λ:

$$h(t) = \lambda$$

Recalling $S(t) = \exp\left[-\int_0^t h(u)du\right]$, the survivor function is

$$S(t) = \exp\left(-\int_0^t \lambda\right) = \exp(-\lambda t)$$

What is the CDF of the survival times?

$$\begin{aligned} P(T < t) &\equiv 1 - P(T \geq t) \\ &\equiv 1 - S(t) = 1 - \exp(-\lambda t) \end{aligned}$$

The distribution of survival times is *exponential*(λ)!

Fundamental goals of survival analysis

1. Estimate survivor and/or hazard functions from data $\Longrightarrow$ estimator will depend on your choice of model (what assumptions you impose).
2. Compare survival and hazard functions between groups $\Longrightarrow$ Hypothesis testing!
3. Assess the relationship of explanatory variables to survival time $\Longrightarrow$ Regression modelling!

Summary

1. Survival data is the time until the occurrence of an event of interest.
2. Survival data is almost always subject to incompleteness – right censoring is the most common but other forms abound.
3. Survival analysis methods must account for the presence of incomplete information to (a) make efficient use of the available data and (b) avoid bias in estimates of useful parameters.
4. The survivor function and hazard functions are two distinct but related quantities that are central to most survival analysis methods.