Biostatistics (MATH11230)

Introduction to survival analysis

Vanda Inácio

University of Edinburgh

Semester 1, 2022/2023

Introduction to survival analysis

General context

- For most of the methodological exposition, we follow Collett (2014, chapter 1).
- In medical studies, a common outcome is the time until an event occurs. Examples include:
  - The time until a patient dies.
  - The time until the recurrence of a cancer that is in remission.
  - The time until a kidney transplant patient needs a new kidney.
- Data involving such an outcome is called '**time-to-event**' data and the branch of statistics that deals with analysing these data is called **survival analysis**.
- Time to event data is a fundamental different type of data and hence the need for a specialised branch of statistics to deal with it.

General context

- Throughout, I will interchangeably use *survival time* and *event time*.
- We shall note that survival analysis methods can also be used beyond the medical research context, such as:
  - Survival times of animals in an experimental study.
  - The time taken by an individual to complete a task in a psychological experiment.
  - The lifetime of industrial or electronic components (also known as reliability).

Introduction to survival analysis

Special features of survival data

- ↪ We must consider the reasons why survival data cannot be analysed with standard statistical procedures.

- ↪ One reason (but not the major one) is that survival data are usually not symmetrically distributed. Note that survival times are always positive.

- ↪ Typically, a histogram constructed from the event times of a group of similar individuals will tend to be skewed to the right, that is, the histogram will have a long tail to the right.

- ↪ For these reasons, it will not be reasonable to assume that event times data can be modelled with the popular normal distribution.

Introduction to survival analysis

Special features of survival data

- ↪ This difficulty could be, of course, overcome by applying a transformation to the data, e.g., the logarithmic one, in order to obtain a more symmetric distribution.
- ↪ However, a more satisfactory approach is to adopt an alternative distribution for modelling the data in its original scale.
- ↪ Popular distributions for modelling event time data are the lognormal, gamma, and weibull distributions (distributions whose support is $\mathbb{R}^+$).

Introduction to survival analysis

Special features of survival data

- ↪ The main feature of survival data that renders traditional statistical methods inappropriate is that event times are frequently **censored**.
- ↪ The event time of an individual is said to be censored when the event of interest has not been observed for that individual.
- ↪ This may be because at the end of the study period the event of interest (death, cancer recurrence, etc) has not occurred.
- ↪ For example, in an hypothetical example of time until patients suffer an heart attack, some patients will never experience an heart attack.
- ↪ Even if the study is about an event that is certain to eventually occur, such as death, it is typically infeasible (due to cost and logistic reasons) to run the study indefinitely until the event occurs.
- ↪ For example, in the hypothetical example about the time until patients who had a kidney transplant need a new kidney, this time can be up to 20 years.

Introduction to survival analysis

Special features of survival data

- ↪ Another reason why an event time may be censored is due to losses to follow up or dropouts.
- ↪ As an example, suppose that after being recruited for a clinical trial, a patient moves to another part of the country, or even to a different country, and can no longer be traced.
- ↪ The only information available about the event time of that patient is the last date on which he/she was known to not have experienced the event.
- ↪ This date may well be the last time that the patient visited the clinic.
- ↪ Another cause of censoring are competing risks. For example, if one is interested in studying the time until cancer recurrence after treatment, a patient may die from another cause before one gets to see when their cancer could have come back.

Introduction to survival analysis

Special features of survival data

- ↪ There are three types of censoring: **right**, **interval**, and **left** censoring.
- ↪ The examples given in the previous slides are all examples of right censoring.
- ↪ So, in right censoring, the endpoint occurs after the observed survival time.
- ↪ For example, subjects enter into a trial at time 0, and the time until death is measured. At time t^* the trial ends but some subjects are still alive. We know that they have survived until time t^*, but we do not know how much longer.
- ↪ Survival time is said to be (right) censored at time t^*.
- ↪ The right censored survival time t^* is less that the actual/true, but unknown, survival time.

Introduction to survival analysis

Special features of survival data

↪ In interval censoring we do know the exact event time, but we know it falls in some interval.

↪ For instance, interval censoring arises if we survey patients once per week in order to determine whether the event has occurred (e.g., onset of symptoms of some disease) and the event happens between two surveys.

↪ All that is therefore known is that the event has happened between such two surveys.

Introduction to survival analysis

Special features of survival data

- ↪ In left censoring, the actual/true survival time is less than that observed.
- ↪ As an example, consider a study in which interest focuses on time to recurrence of a particular cancer following surgical removal of the primary tumour.
- ↪ Three months after their operation, the patients are examined to determine if the cancer has recurred.
- ↪ At such time, the can cancer has already recurred for some patients. That is, for these patients, the actual time to recurrence is less than three months.
- ↪ Left censored data also commonly arise when measurement instruments are inaccurate below a lower limit of detection and, as such, this limit is then reported.

Special features of survival data

↪ Any meaningful analysis of time to event data has to take censoring into account, doing so is what survival analysis is all about!

↪ The intuitive approaches of: (i) treating the censored times as the actual/true event times, and (ii) discarding the censored observations lead, most of the times, to erroneous conclusions.

↪ In this course, we will focus on right censoring. In some cases, adaptations of the methods we will learn to deal with right censored data can be made so that the method is also applicable to left and interval censoring, but that is beyond the scope of this course.

Introduction to survival analysis

Special features of survival data

- An important assumption that will be made is that, conditional on relevant prognostic variables, the event time is independent of the censoring time.

- For instance, suppose that some patients dropout of a cancer study early because they are already very ill. This clearly violates the independence assumption.

- Typically it is not possible to determine from the data itself whether the censoring mechanism is independent.

- Instead, one has to carefully consider the data collection process in order to determine whether independent censoring is a reasonable assumption.

Introduction to survival analysis

Survival, hazard, and cumulative hazard functions

- Let T be a nonnegative and continuous random variable representing the time until an event and let t be an actual survival time.
- The distribution of T can be specified in a variety of ways, through:
  - The survival function.
  - The hazard function.
  - The cumulative hazard function.

Introduction to survival analysis

Survival, hazard, and cumulative hazard functions

- ↪ Let the probability density function of T be given by $f(t)$.

- ↪ The (cumulative) distribution function of T is then given by

$$F(t) = \Pr(T \leq t) = \int_0^t f(u)\mathrm{d}u, \quad t > 0.$$

- ↪ The distribution function represents the probability that the survival time is less or equal than some value t.

- ↪ Obviously, it also holds that

$$f(t) = \frac{\mathrm{d}}{\mathrm{d}t}F(t).$$

Introduction to survival analysis

Survival, hazard, and cumulative hazard functions

- The survival function, denoted by $S(t)$, is defined to be the probability that the survival time exceeds t, and we have that

$$S(t) = \Pr(T > t) = \int_t^{\infty} f(u)\mathrm{d}u = 1 - F(t).$$

- The survival function can therefore be used to represent the probability that an individual survives beyond any given time.

- Trivially, we also have that

$$f(t) = \frac{\mathrm{d}}{\mathrm{d}t}F(t) = \frac{\mathrm{d}}{\mathrm{d}t}(1 - S(t)) = -\frac{\mathrm{d}}{\mathrm{d}t}S(t).$$

Introduction to survival analysis

Survival, hazard, and cumulative hazard functions

- ↪ The hazard function is widely used to express the risk of an event (e.g., death) occurring at some time t.

- ↪ The hazard function is obtained from the probability that the event occurs at time t, conditional on the event not having occurred before t.

- ↪ For a formal definition of the hazard function, consider the probability that the random variable representing the time until the event, lies between t and $t + \Delta t$, conditional on T being greater or equal than t, written as $\Pr(t \leq T < t + \Delta t \mid T \geq t)$.

- ↪ This conditional probability is then expressed as a probability per unit time by dividing by the time interval Δt to give a rate.

Introduction to survival analysis

Survival, hazard, and cumulative hazard functions

↪ The hazard function, $h(t)$ is then the limiting value of this quantity, as Δt tends to zero, so that

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{\Pr(\{t \leq T < t + \Delta t\} \cap \{T \geq t\})}{\Pr(T \geq t)\Delta t} \\ &= \frac{1}{S(t)} \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t)}{\Delta t} \\ &= \frac{1}{S(t)} \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \\ &= \frac{f(t)}{S(t)}. \end{aligned}$$

↪ Note that

$$\lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t}$$

is the definition of the derivative of $F(t)$ with respect to t, which is $f(t)$.

- ↪ For instance, if the survival time is measured in days, $h(t)$ is the approximate probability that an individual, who is at risk of the event occurring at the start of day t, experiences the event during that day.

- ↪ The function $h(t)$ is also known as the hazard rate, the instantaneous death rate, the intensity rate, or the force of mortality.

- ↪ As its definition makes clear, the hazard function is nonnegative.

- ↪ As the survival (or distribution) function, the hazard function uniquely defines the distribution of T (under the assumption that $f(t)$ is continuous).

↪ An alternative expression for the hazard function is given by

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} \\ &= \frac{-\frac{\mathrm{d}}{\mathrm{d}t} S(t)}{S(t)} \\ &= -\frac{\mathrm{d}}{\mathrm{d}t} \log S(t). \end{aligned}$$

Introduction to survival analysis

Survival, hazard, and cumulative hazard functions

- Finally, the cumulative hazard function, $H(t)$, is simply the cumulative risk of an event occurring by time t:

$$H(t) = \int_0^t h(u) \mathrm{d}u.$$

- If the event is death, then $H(t)$ summarises the risk of death up to time t, given that death has not occurred before t.

- Note that $H(t)$ can also be written as

$$\begin{aligned} H(t) &= \int_0^t h(u) \mathrm{d}u \\ &= \int_0^t -\frac{\mathrm{d}}{\mathrm{d}u} \log S(u) \mathrm{d}u \\ &= -[\log S(u)]_0^t \\ &= -\log S(t) + \log S(0), \qquad S(0) = \Pr(T > 0) = 1 \\ &= -\log S(t). \end{aligned}$$

Introduction to survival analysis

Survival, hazard, and cumulative hazard functions

- ↪ The equality $H(t) = -\log S(t)$, also implies that

$$S(t) = \exp\{-H(t)\}.$$

- ↪ Further, in some books about survival analysis you may find the density function expressed as

$$f(t) = h(t)\exp\{-H(t)\}.$$

- ↪ This comes from the fact that we know that

$$h(t) = \frac{f(t)}{S(t)}, \qquad S(t) = \exp\{-H(t)\}.$$

- ↪ Once again, $H(t)$ uniquely defines the distribution of T.

- The simplest survival distribution is the exponential distribution, whose density function is given by
$$f(t) = \lambda e^{-\lambda t}, \quad \lambda > 0.$$
- The corresponding distribution and survival functions are defined as
$$F(t) = 1 - e^{-\lambda t}, \quad S(t) = e^{-\lambda t}.$$
- The hazard function associated with the exponential distribution is given by
$$h(t) = \frac{f(t)}{S(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda.$$
- The fact that the hazard function is constant over time, may be unrealistic for most situations, and this is the reason why the exponential distribution is not used often as a distribution for event/survival data.

Introduction to survival analysis

Survival, hazard, and cumulative hazard functions

- The lognormal distribution is often used to model survival data. Its density function is given by

$$f(t) = \frac{1}{t\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log t - \mu)^2}{2\sigma^2}\right), \quad \mu \in \mathbb{R}, \quad \sigma > 0.$$

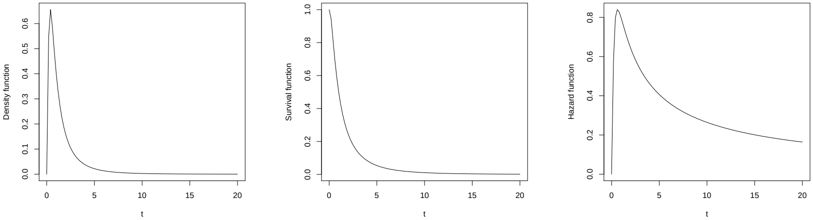
- The corresponding distribution function is not pleasing, and as a consequence, the survival, hazard, or cumulative hazard functions are also not pleasing to work with from an analytical point of view.

- However, and fortunately, we can use the functions `dlnorm` and `plnorm` in `R`.

Introduction to survival analysis

Survival, hazard, and cumulative hazard functions

- Below, the density, survival, and hazard functions of the lognormal distribution with $\mu = 0$ and $\sigma = 1$.


Density function
t
Survival function
t
Hazard function
t