Survival analysis

Christiana Kartsonaki

Abstract

Survival analysis is the analysis of data involving times to some event of interest. The distinguishing features of survival, or time-to-event, data and the objectives of survival analysis are described. Some fundamental concepts of survival analysis are introduced and commonly used methods of analysis are described.

Keywords Cox proportional hazards model; failure times; hazard; Kaplan–Meier curve; survival data; time-to-event data

Introduction

Survival analysis is the analysis of time-to-event data. Such data describe the length of time from a time origin to an endpoint of interest. For example, individuals might be followed from birth to the onset of some disease, or the survival time after the diagnosis of some disease might be studied. Survival analysis methods are usually used to analyse data collected prospectively in time, such as data from a prospective cohort study or data collected for a clinical trial.

The time origin must be specified such that individuals are as much as possible on an equal footing. For example if the survival time of patients with a particular type of cancer is being studied, the time origin could be chosen to be the time point of diagnosis of that type of cancer. Equally importantly, the endpoint or event of interest should be appropriately specified, such that the times considered are well-defined. In the above example, this could be death due to the cancer studied. Then the length of time from the time origin to the endpoint could be calculated.

One of the reasons why survival analysis requires 'special' techniques is the possibility of not observing the event of interest for some individuals. For example individuals may drop out of a study, or they might have a different event, such as in the above example death due to an accident, which is not part of the endpoint of interest. Another possibility is that there might be a time point at which the study finishes and thus if any individuals have not had their event yet, their event time will not have been observed. These incomplete observations cannot be ignored, but need to be handled differently. This is called *censoring*. Another feature of survival data is that distributions are often skewed (asymmetric) and thus simple techniques based on the normal distribution cannot be directly used.

The objectives of survival analysis include the analysis of patterns of event times, the comparison of distributions of survival times in different groups of individuals and examining whether and by how much some factors affect the risk of an event of interest.

Christiana Kartsonaki *DPhil Nuffield Department of Population Health, University of Oxford, Oxford, UK. Conflict of interest statement: none.*

Censoring

The most commonly encountered type of censoring and easiest to handle in the analysis is *right censoring*. Right censoring occurs when an individual is followed up from a time origin t_0 up to some later time point t_C and he/she has not had the event of interest, such that all we know is that their event has not occurred up to their censoring time t_C. This may occur, for example, if an individual drops out of a study before the event of interest occurs. Commonly studies are terminated at some specified time and at the end of the study some individuals have not yet had their event. This is sometimes referred to as *administrative* censoring. In some studies the majority of participants are censored. Event and censoring times of 10 patients are illustrated in Figure 1.

Another type of censoring is *left censoring*. Left censoring is the situation in which an individual is known to have had the event before a specific time, but that could be any time before the censoring time. It is also possible to have *interval censoring* where an individual is only known to have had the event between two time points but the exact time of event is not observed.

A different concept is *truncation*. Truncation is something that happens by design. *Left truncation* is the most commonly encountered type of truncation, where individuals enter the study after they have their truncation event (which is not the same as the event being studied). Delayed entry where for example a set of adults are recruited into a study but those who had the event before adulthood are not included at all is very common. *Right truncation* occurs when the entire study population has already experienced the event of interest.

For the standard methods of analysis that we focus on here censoring should be *non-informative*, that is, the time of censoring should be independent of the event time that would have otherwise been observed, given any explanatory variables included in the analysis, otherwise inference will be biased.

An example of informative censoring which must not be ignored is as follows: in a study of survival after a disease diagnosis, patients might be lost to follow up because their condition has become worse and are no longer able to attend appointments. Or in a study of treatments for a non-life-threatening condition, some patients might drop out of the study because their condition has improved and they choose to discontinue treatment. It is usually not possible to know whether the censoring in a study really is non-informative.

Example. Data from a clinical trial on colon cancer adjuvant therapy[1] are used as an illustration. A group of colon cancer patients are followed up from diagnosis to death. That is, the *time scale* has origin the time of diagnosis of colon cancer and endpoint the time of death from colon cancer. The dataset, freely available in the statistical software R[2] (dataset 'colon' in package 'survival'[3]), contains observations on 929 colon cancer patients. These are the first 10 observations on a subset of the variables:

id	status	time	sex	age	nodes	differ	surg	node4
1	1	1521	1	43	5	2	0	1
2	0	3087	1	63	1	2	0	0
3	1	963	0	71	7	2	0	1
4	1	293	0	66	6	2	1	1
5	1	659	1	69	22	2	1	1
6	1	1767	0	57	9	2	0	1
7	1	420	1	77	5	2	1	1
8	0	3192	1	54	1	2	0	0
9	0	3173	1	46	2	2	0	0
10	0	3308	0	68	1	2	1	0

The variable 'status' indicates whether a patient has died, taking the value 1 if a patient has died and 0 otherwise, and 'time' is the survival time since diagnosis in days. 'age' is the patients' age at the time of entry into the study, 'nodes' is the number of lymph nodes with detectable cancer and 'node4' is a binary variable taking the value 1 if the patient has more than four lymph nodes with cancer and 0 if the patient has fewer than or equal to four positive lymph nodes. The event and censoring times are illustrated in Figure 1.

Some definitions

Let $T \geq 0$ be a random variable representing the survival (or event) time. The *survival* (or *survivor*) *function* is the probability that an individual survives beyond time t,

$$S(t) = \mathbb{P}(T > t), \ 0 < t < \infty.$$

The *probability density function* $f(t)$ is the frequency of events per unit time. The probability density function is related to the survival function,

$$f(t) = -\frac{dS(t)}{dt}.$$

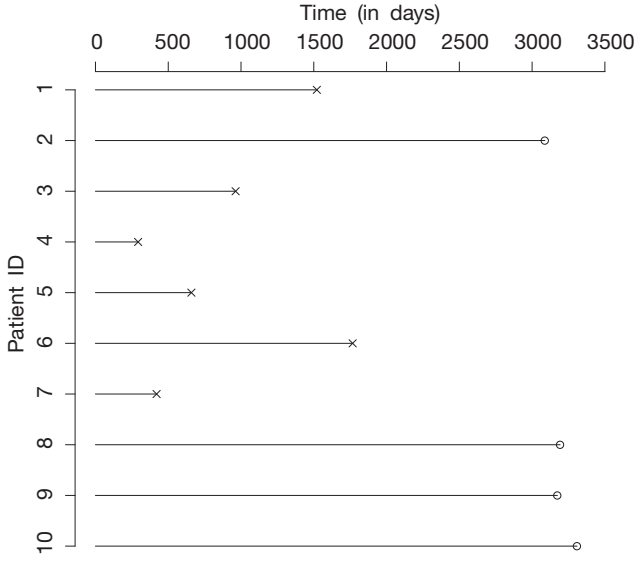

Figure 1 Survival times (×) and censoring times (○) in days.

The *hazard function* is the instantaneous rate at which events occur for individuals which are surviving at time t,

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

and the *cumulative hazard function* is

$$H(t) = \int_0^t h(u)du.$$

The cumulative hazard function is related to the survival function as follows:

$$S(t) = e^{-H(t)}.$$

That is, the higher the hazard, the lower the survival.

Let δ_i be equal to 1 for individual i if individual i had the event and 0 if individual i was censored. Then for a set of possibly right-censored data, the data for individual i can be represented as (t_i, δ_i, x_i), where t_i is the time of event or censoring, δ_i is a censoring indicator and x_i are the covariates, that is, a set of variables representing any other information on that individual.

Then the *likelihood function* is

$$L = \prod_{j \text{ had event}} f(t_j) \prod_{k \text{ censored}} S(t_k) = \prod_{i=1}^{N} h(t_i)^{\delta_i} S(t_i).$$

That is, each individual with an observed event time t_i contributes the hazard rate at t_i multiplied by the survival to t_i and each individual that is censored at t_i contributes the survival to t_i.

Estimation

One objective of the analysis of time-to-event data is given a set of data to estimate and plot the survival function.

A very widely used method of doing that is calculating and plotting a Kaplan–Meier curve. This is a non-parametric method of estimating the survival function. Non-parametric methods are rather simple methods which do not make any distributional assumptions, in this context about the distribution of survival times observed in a study. Non-parametric methods are very useful for summarizing survival data and making simple comparisons but cannot so easily deal with more complex situations.

Let $t_1 < t_2 < \ldots < t_k$ be the observed event times and $n = n_0$ the sample size. Let d_j be the number of individuals who have an event at time t_j, where $j = 1, \ldots, k$, and m_j the number of individuals censored in the interval $[t_j, t_{j+1})$. Then $n_j = (m_j + d_j) + \ldots + (m_k + d_k)$ is the number of individuals at risk just prior to t_j.

The *Kaplan–Meier* (or *product-limit*) *estimator*[4] is a non-parametric estimator of the survival function,

$$\widehat{S}(t) = \prod_{j:\, t_j \leq t} \frac{n_j - d_j}{n_j}.$$

Standard errors can be calculated using *Greenwood's formula*,[5] which approximates the variance as