

Lecture 11

Survival analysis: Introduction

Reading: Cox and Oakes, chapter 1, 4, section 11.6, CT4 Unit 6.3-6.5, Klein and Moeschberger sections 3.1–3.4, Chapter 4

11.1 Incomplete observations: Censoring and truncation

We begin by considering simple analyses but we will lead up to and take a look at regression on explanatory factors, as in linear regression part A. The important difference between survival analysis and other statistical analyses which you have so far encountered is the presence of censoring. This actually renders the survival function of more importance in writing down the models.

Right censoring occurs when a subject leaves the study before an event occurs, or the study ends before the event has occurred. For example, we consider patients in a clinical trial to study the effect of treatments on stroke occurrence. The study ends after 5 years. Those patients who have had no strokes by the end of the year are censored. If the patient leaves the study at time t_e , then the event occurs in (t_e, ∞) .

Left censoring is when the event of interest has already occurred before enrolment. This is very rarely encountered.

Truncation is deliberate and due to study design.

Right truncation occurs when the entire study population has already experienced the event of interest (for example: a historical survey of patients on a cancer registry).

Left truncation occurs when the subjects have been at risk before entering the study (for example: life insurance policy holders where the study starts on a fixed date, event of interest is age at death).

Generally we deal with **right censoring & sometimes left truncation**.

Two types of independent **right censoring**:

Type I : completely random dropout (eg emigration) and/or fixed time of end of study no event having occurred.

Type II: study ends when a fixed number of events amongst the subjects has occurred.

Skeptical question: Why do we need special techniques to cope with incomplete observations? Aren't all observations incomplete? After all, we never see all possible samples from the distribution. If we did, we wouldn't need any sophisticated statistical analysis.

The point is that most of the basic techniques that you have learned assume that the observed values are interchangeable with the unobserved values. The fact that a value has been observed does not tell us anything about what the value is. In the case of censoring or

truncation, there is dependence between the event of observation and the value that is observed. In right-censoring, for instance, the fact of observing a time implies that it occurred *before* the censoring time. The distribution of a time *conditioned on its being observed* is thus different from the distribution of the times that were censored.

There are different levels of independence, of course. In the case of Type I censoring, the censoring time itself **is** independent of the (potentially) observed time. In Type II censoring, the censoring time depends in a complicated way on all the observation times.

11.2 Likelihood and Censoring

If the censoring mechanism is *independent* of the event process, then we have an easy way of dealing with it.

Suppose that T is the time to event and that C is the time to the censoring event.

Assume that all subjects may have an event or be censored, say for subject i one of a pair of observations $(\tilde{t}_i, \tilde{c}_i)$ may be observed. Then since we observe the minimum time we would have the following expression for the likelihood (using independence)

$$L = \prod_{\tilde{t}_i < \tilde{c}_i} f(\tilde{t}_i) S_C(\tilde{t}_i) \prod_{\tilde{c}_i < \tilde{t}_i} S(\tilde{c}_i) f_C(\tilde{c}_i)$$

Now define the following random variable:

$$\delta = \begin{cases} 1 & \text{if } T < C \\ 0 & \text{if } T > C \end{cases}$$

For each subject we observe $t_i = \min(\tilde{t}_i, \tilde{c}_i)$ and δ_i , observations from a continuous random variable and a binary random variable. In terms of these L becomes

$$L = \prod_i h(t_i)^{\delta_i} S(t_i) \prod_i h_C(t_i)^{1-\delta_i} S_C(t_i)$$

where we have used density = hazard \times survival function.

NB If the censoring mechanism is independent (sometimes called non-informative) then we can ignore the second product on the right as it gives us no information about the event time. In the remainder of the course we will assume that the censoring mechanism is independent.

11.3 Data

Demographic v. trial data

Our models include a “time” parameter, whose interpretation can vary. First of all, in population-level models (for instance, a birth-death model of population growth, where the state represents the number of individuals) the time is true calendar time, while in individual-level models (such as our multiple-decrement model of death due to competing risks, or the healthy-sick-dead process, where there is a single model run for each individual) the time parameter is more likely to represent individual age. Within the individual category, the time to event can literally be the age, for instance in a life insurance policy. In a clinical trial it will more typically be time from admission to the trial.

For example, consider the following data from a Sydney hospital pilot study, concerning the treatment of bladder cancer:

Time to cancer	Time to recurrence	Time between	Recurrence status
0.000	4.967	4.967	1
21.020	22.993	1.974	1
45.033	61.086	16.053	0
52.171	55.033	2.862	1
48.059	65.033	16.974	0

All times are in months. Each patient has their own zero time, the time at which the patient entered the study (accrual time). For each patient we record time to event of interest or censoring time, whichever is the smaller, and the status, $\delta = 1$ if the event occurs and $\delta = 0$ if the patient is censored. If it is the recurrence that is of interest, so in fact the relevant time, the “time between”, is measured relative to the zero time that is the onset of cancer.

11.4 Non-parametric survival estimation

11.4.1 Review of basic concepts

Consider random variables X_1, \dots, X_n which represent independent observations from a distribution with cdf F . Given a class \mathfrak{F} of possibilities for F , an **estimator** is a choice of the “best”, on the basis of the data. That is, it is a function from \mathbb{R}_+^n to \mathfrak{F} which maps an collection of observations (x_1, \dots, x_n) to \mathfrak{F} .

Estimators for distribution functions may be either **parametric** or **non-parametric**, depending on the nature of the class \mathfrak{F} . The distinction is not always clear-cut. A parametric estimator is one for which the class \mathfrak{F} depends on some collection of parameters. For example, it might be the two-dimensional family of all gamma distributions. A non-parametric estimator is one that does not impose any such parametric assumptions, but allows the data to “speak for themselves”. There are intermediate non-parametric approaches as well, where an element of \mathfrak{F} is not defined by any small number of parameters, but is still subject to some constraint. For example, \mathfrak{F} might be the class of distributions with smooth hazard rate, or it might be the class of log-concave distribution functions (equivalent to having increasing hazard rate). We will also be concerned with **semi-parametric** estimators, where an underlying infinite-dimensional class of distributions is modified by one or two parameters of special interest.

The disadvantage of parametrisation is always that it distorts the observations; the advantage is that it allows the data from different observations to be combined into a single parameter estimate. (Of course, if the data are *known* to come from some distribution in the parametric family, the “distortion” is also an advantage, because the real distortion was in the data, due to random sampling.)

We start by considering nonparametric estimators of the cdf. These have the advantage of limiting the assumptions imposed upon the data, but the disadvantage of being too strictly limited by the data. That is, taken literally, the estimator we obtain from a sample of observed times will imply that only exactly those times actually observed are possible.

If there are observations x_1, \dots, x_n from a random sample then we define the empirical distribution function

$$\hat{F}(x) = \frac{1}{n} \# \{x_i : x_i \leq x\}$$

This is an appropriate non-parametric estimator for the cdf if no censoring occurs. However if censoring occurs this has to be taken into account.

We measure the pair (X, δ) where $X = \min(T, C)$ and δ is as before

$$\delta = \begin{cases} 1 & \text{if } T < C \\ 0 & \text{if } T > C \end{cases}$$

Suppose that the observations are (x_i, δ_i) for $i = 1, 2, \dots, n$.

$$\begin{aligned} L &= \prod_i f(x_i)^{\delta_i} S(x_i)^{1-\delta_i} \\ &= \prod_i f(x_i)^{\delta_i} (1 - F(x_i))^{1-\delta_i} \end{aligned}$$

What follows is a heuristic argument allowing us to find an estimator for S , the survival function, which in the likelihood sense is the best that we can do. Notice first that there is no MLE if we model the failure time as a continuous random variable. Suppose T has density f , with survival function $S = 1 - F$

Suppose that there are failure times $(t_0 = 0 <) t_1 < \dots < t_i < \dots$. Let $s_{i1}, s_{i2}, \dots, s_{ic_i}$ be the censoring times within the interval $[t_i, t_{i+1})$ and suppose that there are d_i failures at time t_i (allowing for tied failure times). Then the likelihood function becomes

$$\begin{aligned} L &= \prod_{fail} f(t_i)^{d_i} \prod_i \left(\prod_{k=1}^{c_i} (1 - F(s_{ik})) \right) \\ &= \prod_{fail} (F(t_i) - F(t_{i-}))^{d_i} \prod_i \left(\prod_{k=1}^{c_i} (1 - F(s_{ik})) \right) \end{aligned}$$

where we write $f(t_i) = F(t_i) - F(t_{i-})$, the difference in the cdf at time t_i and the cdf immediately before it.

Since $F(t_i)$ is an increasing function, and *assuming that it takes fixed values at the failure time points*, we make $F(t_{i-})$ and $F(s_{ik})$ as small as possible in order to maximise the likelihood. That means we take $F(t_{i-}) = F(t_{i-1})$ and $F(s_{ik}) = F(t_i)$.

This maximises L by considering the cdf $F(t)$ to be a step function and therefore to come from a discrete distribution, with failure times as the actual failure times which occur. Then

$$L = \prod_{fail} (F(t_i) - F(t_{i-1}))^{d_i} \prod_i (1 - F(t_i))^{c_i}$$

So we have showed that amongst all cdf's with fixed values $F(t_i)$ at the failure times t_i , then the discrete cdf has the maximum likelihood, amongst those with d_i failures at t_i and c_i censorings in the interval $[t_i, t_{i+1})$.

Let us consider the **discrete case** and let

$$P\{\text{fail at } t_i | \text{survived to } t_{i-}\} = h_i$$

Then

$$\begin{aligned} S(t_i) &= 1 - F(t_i) = \prod_1^i (1 - h_j), \\ f(t_i) &= h_i \prod_1^{i-1} (1 - h_j) \end{aligned}$$

Finally we have

$$L = \prod_{t_i} h_i^{d_i} (1 - h_i)^{n_i - d_i}$$

where n_i is the number at risk at time t_i . This is usually referred to as the number in the risk set.

Note

$$n_{i+1} + c_i + d_i = n_i$$

11.4.2 Kaplan-Meier estimator

This estimator for $S(t)$ uses the mle estimators for h_i . Taking logs

$$l = \sum_i d_i \log h_i + \sum_i (n_i - d_i) \log(1 - h_i)$$

Differentiate with respect to h_i

$$\begin{aligned} \frac{\partial l}{\partial h_i} &= \frac{d_i}{h_i} - \frac{n_i - d_i}{1 - h_i} = 0 \\ \implies \hat{h}_i &= \frac{d_i}{n_i} \end{aligned}$$

So the Kaplan-Meier estimator is

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

where

$$\begin{aligned} n_i &= \#\{\text{in risk set at } t_i\}, \\ d_i &= \#\{\text{events at } t_i\}. \end{aligned}$$

Note that $c_i = \#\{\text{censored in } [t_i, t_{i+1})\}$. If there are no censored observations before the first failure time then $n_0 = n_1 = \#\{\text{in study}\}$. Generally we assume $t_0 = 0$.

11.4.3 Nelson-Aalen estimator and new estimator of S

The Nelson-Aalen estimator for the cumulative hazard function is

$$\hat{H}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i} \quad \left(= \sum_{t_i \leq t} \hat{h}_i \right)$$

This is natural for a discrete estimator, as we have simply summed the estimates of the hazards at each time, instead of integrating, to get the cumulative hazard. This correspondingly gives an estimator of S of the form

$$\begin{aligned} \tilde{S}(t) &= \exp(-\hat{H}(t)) \\ &= \exp\left(-\sum_{t_i \leq t} \frac{d_i}{n_i}\right) \end{aligned}$$

It is not difficult to show by comparing the functions $1-x$, $\exp(-x)$ on the interval $0 \leq x \leq 1$, that $\tilde{S}(t) \geq \hat{S}(t)$.

11.4.4 Invented data set

Suppose that we have 10 observations in the data set with failure times as follows:

$$2, 5, 5, 6+, 7, 7+, 12, 14+, 14+, 14+ \quad (1)$$

Here + indicates a censored observation. Then we can calculate both estimators for $S(t)$ at all time points. It is considered unsafe to extrapolate much beyond the last time point, 14, even with a large data set.

Table 11.1: Computations of survival estimates for invented data set (1)

t_i	d_i	n_i	\hat{h}_i	$\hat{S}(t_i)$	$\tilde{S}(t_i)$
2	1	10	0.10	0.90	0.90
5	2	9	0.22	0.70	0.72
7	1	7	0.14	0.60	0.63
12	1	4	0.25	0.45	0.54