Duration (survival) data in healthcare

Tool: Kaplan-Meier estimator, Cox proportional hazard model.

The Analytics Edge: Survival data are common in different domains, such as medical research or transportation engineering. They generally describe situations in which an event of interest (e.g., the death of a patient or the failure of a mechanical component) can happen at different times. Descriptive analytics can help estimate the survival rate of patients or predict the timing of failures.

Overview

Duration (or survival) analysis aims at analyzing the expected duration of time until a specific *event* happens. In other words, the *duration* is the time from the beginning of an observation period to the event of interest (or the end of the study or the loss of contact/withdrawal from the study). Consider the example illustrated in Figure 0.1: at time t, we observe the status of a patient after a heart transplant. At that time, the patient might be dead or alive. If the patient is alive (so he/she does not have an event during the observation time), we are dealing with a *censored* observation (the subject is censored in the sense that nothing is observed or known about that subject after the time of censoring). So, the survival time for the alive patient is at least t, but not exactly t. Our interest is to describe the *survival function* $S(t)$, which expresses the probability that a subject survives longer than time t.

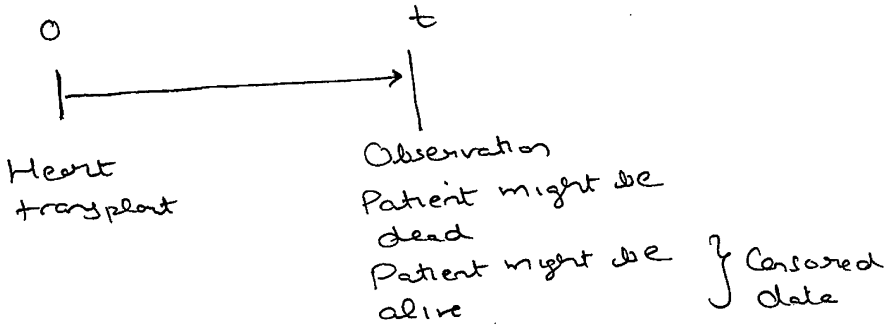

Figure 0.1: Survival data for patients after a surgery.

Another example of duration analysis is a government studying the length of unemployment for a group of citizens. In such case, the event of interest could be the time at which a citizen finds a job. Duration analysis finds application in several sectors, such as transport engineering, where analysts are typically interested in estimating the length of time until failure of a certain mechanical component. Note that duration times are by definition nonnegative.

Dealing with survival data

Survival function and hazard rate

Let's denote with T a continuous random variable that represents the waiting time until the occurrence of an event (e.g., death). T has a probability density function $f(t)$ and cumulative density function $F(t) = P(T < t)$ (giving the probability that the event has occurred by duration t). The *survival function* $S(t)$ can thus be defined as

$$S(t) = P(T \geq t) = 1 - F(t).$$

The survival function represents the probability of being alive just before duration t, that is, the probability that the event of interest has not occurred by duration t.

One important question in analysis of duration data is: given that an event has lasted until t, what is the probability that it will end in the shortest interval of time Δt? This probability is given by:

$$P(t \leq T \leq t + \Delta t).$$

We can now introduce the *hazard rate* $\lambda(t)$, namely the instantaneous rate of occurrence of the event, which is defined as:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t}.$$

The numerator of this expression is the conditional probability that the event will occur in the interval $[t + \Delta t]$, given that it has not occurred before, and the denominator is the width of the interval. Dividing one by the other we obtain a rate of event occurrence per unit of time. Taking the limit as the width of the interval goes down to zero, we obtain an instantaneous rate of occurrence. Let's now rewrite the term $P(t \leq T \leq t + \Delta t | T \geq t)$ as the ratio between $P(t \leq T \leq t + \Delta t)$ (probability that T is in the interval $[t, t + \Delta t]$) and $P(T \geq t)$. This yields:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t)}{P(T \geq t)\Delta t},$$

which can finally be rewritten as:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{S(t)\Delta t} = \frac{f(t)}{S(t)}.$$

In words, the rate of occurrence of the event at duration t equals the density of events at t, divided by the probability of surviving to that duration without experiencing the event.

Kaplan-Meier estimator

In 1958, Kaplan and Meier proposed a non-parametric statistic (also known as product limit estimator) to estimate the survival function from lifetime (censored) data. Let's denote with $t_1 \leq t_2 \leq \ldots \leq t_N$ the distinct ordered times of death. Let d_i be the number of deaths at $t(i)$, and let n_i be the number alive just before $t(i)$. This is the number exposed to risk at time $t(i)$. Then, the Kaplan-Meier estimate of the survival function is:

$$\hat{S}(t) = \prod_{i:t_i < t} \Big(1 - \frac{d_i}{n_i}\Big).$$

A heuristic justification of the estimate is as follows. To survive to time t you must first survive to t_1. You must then survive from t_1 to t_2 given that you have already survived to t_1. And so on. Because there

are no deaths between t_{i-1} and t_i, we take the probability of dying between these times to be zero. The conditional probability of dying at t_i given that the subject was alive just before can be estimated by d_i/n_i. The conditional probability of surviving time t_i is the complement $1 - d_i/n_i$. The overall unconditional probability of surviving to t is obtained by multiplying the conditional probabilities for all relevant times up to t.

Example. Consider the following remission times (in weeks) for 21 patients suffering from leukaemia: 1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23. We can calculate t and $\hat{S}(t)$ as follows:

t	$\hat{S}(t)$
$t < 1$	$\frac{21}{21} = 1$
$1 \leq t < 2$	$1 \cdot \frac{21-2}{21} = 0.905$
$2 \leq t < 3$	$1 \cdot 0.905 \cdot \frac{19-2}{19} = 0.81$
$\ldots$	$\ldots$

The result can be synthesized in the Kaplan-Meier curve (see Figure 0.2), a step function with discontinuities at the observed death times.

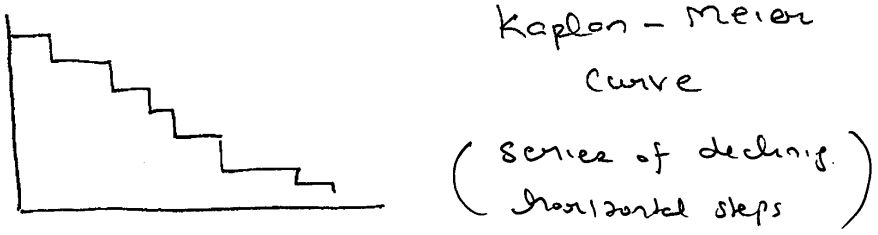

Figure 0.2: Kaplan-Meier curve.

Cox proportional hazard model

The Cox proportional hazard model is a very established model in the bio-medical research domain—in fact, it was also adopted in the Framingham Heart Study. The goal of the model is to describe the hazard rate $\lambda(t)$ as a function of predictors:

$$\lambda(t) = \lambda_0(t) e^{\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p},$$

where t is the survival time, $\{x_1, \ldots, x_p\}$ a set of predictors, $\{\beta_1, \ldots, \beta_p\}$ the corresponding coefficients, and λ_0 the so-called *baseline hazard*, which corresponds to the value of the hazard if all the predictors are

equal to zero. A value of the i-th coefficient $beta_i$ greater than 0 indicates that as the value of the i-th predictor increases, the event hazard increases and thus the length of survival decreases. Put another way, a hazard ratio above 1 indicates a predictor that is positively associated with the event probability, and thus negatively associated with the length of survival.