Survival Analysis

Dr. S. Vishnu Shankar
Senior Research Fellow, Indian Agricultural Statistics Research Institute, New Delhi

ARTICLE ID: 42

Introduction

Survival analysis is a branch of statistics, which deals with the expected duration of time until one or more events occur, such as death in biological organisms and failure in mechanical systems. It is also called as reliability theory or analysis in engineering, duration analysis or duration modelling in economics, and event history analysis in sociology. Survival analysis is generally defined as a set of methods for analysing data where the outcome variable is the time until the occurrence of an event of interest. The event of interest can be death, the occurrence of a disease, marriage, divorce, unemployment after education, etc. In survival analysis, subjects are usually followed over a specified time and the focus is on the time at which the event of interest occurs. Censored or time-to-event data are common in medical research, particularly when studying the survival time of patients. Survival data are also frequently occurring in agricultural research, for example, seed germination times, insect survival in plants, disease occurrence, agricultural insurance, etc. In agricultural research, however, survival analysis has been overlooked because of the availability of analysis tools.

Goals of survival analysis

- Goal 1: To estimate and interpret survivor and/ or hazard functions from survival data.
- Goal 2: To compare survivor and/or hazard functions.
- Goal 3: To assess the relationship of explanatory variables to survival time

Concepts and terminology in survival analysis

Censoring

Censoring is a form of missing data problem in which time to event (Dai et al., 2017) is not observed for reasons, such as termination of study before all recruited subjects have shown the event of interest or the subject has left the study prior to experiencing an event. There are generally three reasons why censoring may occur

- A person does not experience the event before the study ends;
- A person is lost to follow-up during the study period;

- A person withdraws from the study because of death (if death is not the event of interest)

Left Censoring	Interval Censoring	Right Censoring
Time-to-event is **Less** than some value	Time-to-event is **Between** two values	Time-to-event is **Greater** than some value
$t_i < X$	$X_1 < t_i < X_2$	$X < t_i$

Truncation

In censoring, sources/events can be detected, but the values (measurements) are not known completely. But in truncation, the object is detected only if its value is greater than some number and the value is completely known in the case of detection. Truncations occur when the observations are excluded by virtue of their time-to-event. We only know that the value is less than some number. The main difference between censoring and truncation is that censored object is detectable while the object is not even detectable in the case of truncation.

a) **Left Truncation**: An event/source is detected if its measurement is greater than a truncation variable or observations with short time-to-event are excluded. Eg: Neonates with very short survival will likely evade.

b) **Right Truncation**: An event/source is detected if its measurement is less than a truncation variable or observations with long time-to-event are excluded. Eg: Study begins only after the development of disease in the body.

c) **Double Truncation**: This occurs when the time to event of interest in the study sample is in an interval.

Survival Function

It is a function describing the proportion of individuals surviving to or beyond a given time. Let T denote the survival time, then the survival function S (t) is defined as follows

$$S\,(t) = \frac{\textbf{Number of observation survives longer than t}}{\textbf{Total number of observtions}}$$

Hazard function

The hazard function equals the limit, as Δt approaches zero, of a probability statement about survival, divided by Δt, where Δt denotes a small interval of time. The hazard function h(t) gives the instantaneous potential per unit time for the event to occur, given that the individual has survived up to time t.

Hazard Ratio

Exponential of coefficient (e^{Coef}) gives an estimated hazard ratio (HR) for the effect of each variable adjusted for the other variables in a model.

Statistical Methods used in survival analysis

Types	Advantages	Disadvantages	Methods
Parametric	Easy to interpret, efficient and accurate. when the survival times follow a particular distribution.	When the distribution assumption is violated, it may be inconsistent and can give sub-optimal results.	Distribution method Graphical method Regression method
Non-parametric (Highly used)	More efficient when suitable theoretical distributions are not known.	Difficult to interpret.	Kaplan-Meier Nelson-Aalen Life-Table
Semi-parametric (Non-linear)	Used when the relationship between response and some explanatory variable is unknown.	The distribution of the outcome is unknown; not easy to interpret.	Cox model

Kaplan–Meier estimator

Product-limit (PL) method of estimating the survivorship function using right censoring was developed by Kaplan and Meier (1958). The Kaplan–Meier estimator is also known as the product limit estimator. Kaplan Meier's method is suitable when no assumption about the functional distribution of hazard rate.

Nelson-Aalen estimator

The Nelson-Aalen method provides a consistent estimate of the cause-specific hazards. The joint distribution can also be specified by means of the cumulative incidence functions, representing the probability of failing from a given cause before a specific time. Instead of estimating the survival probability as done in the KM estimator, the NA estimator directly estimates the hazard probability. In this case, all causes of failure are involved to estimate the cumulative incidence function of a given cause, and thus other failures cannot be treated as censored observations.

Conclusion

Thus, survival analysis is an age-old statistical technique that can be applied in many areas where our interest is to model the time-to-event occurrence. However, it can be applied to many areas such as the medical field, unemployment studies, crop insurance, insect pest modelling, seed germination, plant growth in long-term experiments, germination test data, botanical epidemiology etc. In general survival analysis can be applied to all the problems where our interest study is time to the happening of an event.