Scope of the Appliaction for Survival Analysis

a_statistician

March 17, 2017

For a study with the survival time or time to event as the random response variable, it is very common that the full information about this kind of response variable is not available, because some subjects are still alive or have no events when the study ends or withdraw from the study before the events happen. Under this situation, the censoring data are generated. It means we know that the events have not happened until the observed time, but we do not know the exact time that events happened on some projects. In another word, for some subjects, we only know that the value of response is larger than a given value. To incorporate the censoring data, many statistical methods were developed.

For example, we have the nonparametric Kaplan-Meier estimator of survival function, semi-parametric Cox proportional hazard model and some parametric models.

Because of the term "survival analysis" and "time to event" response variable, the name "survival analysis" is used to call the statistical methods developed for this kind of data. This name gives statisticians the wrong impression: if response variable is the survival time or time to event, survival analysis must be used; if response variable is not survival time or time to event, the survival analysis is not suitable.

In fact, so called "survival analysis" is the statistical methods developed for incorporating the complicated situation resulted from censoring data by scarifying something as exchange. For example, Kaplan-Meier estimator cannot be used to estimate the mean if the longest censoring time is longer than the observed longest survival time. Now suppose there is no censoring data on the time to event variable. If you use Kaplan-Meier estimator to analyze the data, you need to calculate the area under the estimated survival curve to get the estimate of the mean survival time. In fact, you can estimate the means by sample mean directly. Of course, the survival analysis also scarifies other good properties of the statistical methods used to analyze the data without censoring. In some statistical literature, the non-negativity and/or positive skewness of the response variable were listed as the reasons of using survival analysis. In fact, if there is no censoring data, we have better or equivalent statistical methods corresponding to each survival analysis method. So the censoring data is the only reason that survival analysis methods exist.

So we can say that even the response variable is the survival time or time to event, we should not use the survival analysis given the values of response variable from all subjects are collected.

From other hand, the survival analysis can be used when the response variable is not survival time or time to event, but has censoring data. It means we know that response variable is larger than a given number from some subjects. Here is an example.

Suppose the response variable is the height of the tree. We calculate the height of the tree by measuring the length of the shadow of the tree on the ground on the sunny day and combining other information, such as date, time, longitudinal, latitudinal coordinate, etc. and use geometric knowledge. We can calculate the height of the tree given we measure the length of the shadow correctly and precisely. Suppose some trees in the sample are grown near the river such that the last part of shadow is in the water. Or the shadows of some trees are in the other people's properties. Then we cannot measure the exact length of the shadow (suppose we do not have boat), but we can measure the part of the shadow and claim that shadow is longer than measured length, finally we can say that the tree is higher than converted height. Therefore, we have exact heights on some trees and censored heights on other trees. Under this situation, the survival analysis is the correct methods to use such that the information in the data can be used throughly and the statistical inference is valid. Of course, the assumptions of the survival analysis should be met. For example, censoring is non-informative about the value of response variable; that is, the censoring is caused by something other than the true value of response variable.

In real life, the censoring data from non-time-to-event variable is not rear. Another example is the limit of the measuring instruments, such as PM2.5 for air pollution.

In summary, we should not be confused by the name of survival analysis. Using or not using the survival analysis determined by whether you have censoring data or not in your response variable. (1) If there is no censoring data, other statistical methods are better than survival analysis even if the response variable is the survival time or time to event. (2) Survival analysis is the best methods for response variable with censoring even if it is not the survival time or time to event variable.