Survival Analysis in R

Mark Bounthavong

2/7/2022

This tutorial is located on RPubs.

The entire R Markdown code is located on my GitHub page

For this tutorial, you will need the following packages: `survival`, `dplyr`, `psych`, `survminer`, `gmodels`, and `gtsummary`.

Introduction

A dichotomous variable has two outcomes (Yes or No, Survives or Dies, Cured or Uncured) and is coded a 0 for *No* and 1 for *Yes*. We can compare the differences in the probability or risk of having the event (`outcome = 1`) between two groups using a relative risk ratio or odds ratio. This give us information about the size of the effect, the direction of the effect, and the uncertainty surrounding the effect (e.g., 95% confidence interval). However, we do not have information on how long they were alive or at risk, which is critical information from a survival point of view. Hence, a different kind of analysis was needed to handle this issue–survival analysis.

Survival analysis is also known as a time-to-event analysis. This is a special type of analysis that takes into consideration when the event occurred rather than if the event occurred. In other words, we are focused on acquiring the rate, which is the number of events per unit time.

In survival analysis, we are interested in the hazard, which is the instantaneous event (e.g., death) ro rate at a particular time t.

By combining these two elements (survival and hazard), we will be able to estimate the hazard ratio of an event occurring between two groups.

Survivor and Hazard functions

In survival analysis, the two main features are the survivor and hazard functions.

The survivor function is described as:

Survivor function $S(t)$ is a probability and hazard function $h(t)$ is a rate.

$$S(t) = Pr(T > t),$$

where $S(t)$ is the survival probability, $0 \leq S(t) \leq 1$, and $Pr(T > t)$ is the probability that the time of the event (T) is greater than the some time t.

Hazards function $h(t)$ is the instantaneous potential per unit time for the event to occur, given that the subject is alive (rate)

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

where the hazard rate at some instantaneous point in time $\lim_{\Delta t \to 0}$ is the probability of the event occurring ($P(t \leq T \leq t + \Delta t)$) conditioned on the subject being alive ($T \geq t$ divided by per unit time.

The survivor and hazard functions are related by:

$$h(t) = \lambda \text{ if and only if } S(t) = \exp^{-\lambda t}$$

This means that you can derive the hazard function $h(t)$ from the survivor function $S(t)$ and vice versa.

Kaplan-Meirer Curve

When comparing the survival between two groups, it's good practice to plot