KAPLAN-MEIER SURVIVAL ANALYSIS

Romain Deleris, Grayson Myers

Colorado School of Mines

Problem Definition

A survival curve is defined as

$$S(t) = P(\tau > t)$$

where τ is the survival time and t is a given time. However, the data can be censored at a time c_i before τ_i for a particular case i. The true survival time, τ_i is then unable to be determined for that case.

The Kaplan-Meier estimator seeks then to estimate $S(t)$ even when censoring occurs for some cases.

Assumptions and Simplifications

Kaplan-Meier relies on several assumptions about the data.

- The response variable is binary
- The survival time must be recorded as a specific time rather than an interval, which can be error prone if the discrete-time data has long intervals
- The explanatory variable has minimal left-censoring
- Censoring is independant of the explanatory variable
- There are no cohort effects
- Censoring is consistent across the group

One advantage of Kaplan-Meier, in contrast to other methods of survival analysis is that it can use heavily right-censored data.

However, Kaplan-Meier itself cannot account for confounding variables, so it is often combined with parametric survival analysis techniques if confounding variables are present. [1]

Model

The Kaplan-Meier estimator is given as:

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

where d_i is the number of events (typically deaths) that occurred at time t_i and n_i is the remaining population which has not had an event or been censored.

The data in this case, would be discrete in time, recording the number of events that occurred in that time interval. [2]

Applications

The Kaplan-Meier estimator is applied across various fields within the board context of survival analysis. It is often used in the medical field to assess survival prognosis and treatments due to the frequency of right-censoring in medical studies. It is also used in other fields such as to estimate unemployment duration in economics, subscriber retention in sales, part failure in mechanical engineering, and risk assessment in actuarial science.

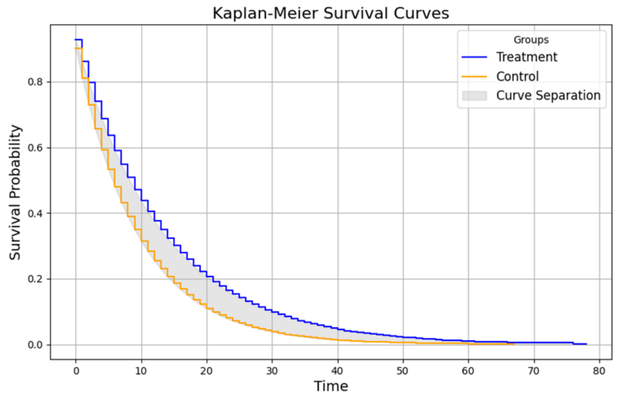
Kaplan-Meier's wide applicability speaks to the versatility of survival analysis and the advantages of Kaplan-Meier specifically.

Simulations

First Simulation :

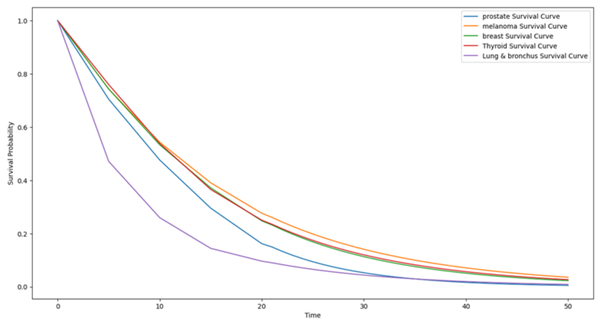
In this case, we use Kaplan-Meier survival analysis to compare the survival curves of a treatment and control group respectively. In industry, a setup like this might be used to estimate the efficacy of a certain treatment over some well-known, competing treatment. These data were generated using an exponential distribution, with the treatment group having rate parameter $\lambda = \frac{1}{5}$ and the control group having rate parameter $\lambda = \frac{1}{4.5}$. Additionally, the probability of censoring is 0.6 for the treatment group and 0.5 for the control group.

In this case, the treatment improves the mean survival time from 4.5 years in the control group to 5 years in the treatment group.



Second Simulation :

In the second simulation, we perform Kaplan-Meier survival analysis on patients presenting with various diseases. This data was sourced from [INSERT SOURCE HERE]. The data included deaths occurring within 5 year intervals from disease onset, which satisfies the assumptions for valid Kaplan-Meier analysis.



These Kaplan-Meier curves can be used to assess survival if a patient presents with a certain disease and provide a baseline to compare to groups receiving various treatments, allowing for informed decisions regarding treatment and realistic expectations for prognosis.

Derivation

By the definition of the survival curve:

$$\begin{aligned} S(t) &= P(\tau > t) \\ &= P(\tau > t \mid \tau > t-1) P(\tau > t-1) \\ &= (1 - P(\tau \leq t \mid \tau > t-1))\, P(\tau > t-1) \\ &= (1 - P(\tau = t \mid \tau \geq t))\, P(\tau > t-1) \\ &= q(t) S(t-1) \end{aligned}$$

where $q(t) = 1 - P(\tau = t \mid \tau \geq t)$

By the recursive equality above, we can say that

$$S(t) = q(t)q(t-1)q(t-2)...q(0)$$

or

$$S(t) = \prod_{i}^{t} q(t-i)$$

The true $q(t)$ cannot be determined if censoring is present, which is why Kaplan-Meier is only an estimator.

Then, let $\tilde{\tau}_k$ be the minimum between the censoring time and the time of the event, or the time of whichever happens first.
For any k such that $c_k < \tau$, we can say that

$$P(\tau = s \mid \tau \geq s) = P(\tilde{\tau}_k = s \mid (\tilde{\tau}_k \geq s)$$

[3]
From there, we can write $\hat{q}(t)$, the estimate of $q(t)$, as:

$$\hat{q}(t) = 1 - \frac{\{1 \leq k \leq n : \tilde{\tau}_k = t\}}{\{1 \leq k \leq n : \tilde{\tau}_k \geq t\}}$$

In this equation, $\{1 \leq k \leq n : \tau_k = s\}$ is the number of events that occur at time t and $\{1 \leq k \leq n : \tau_k = t\}$ is the number of the remaining population which has not experienced an event or been censored. Denoting these quantities as d_k and n_k respectively, we arrive at the definition of the Kaplan-Meier estimator:

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

[3]

References

[1] Michael Foley. *Survival Analysis in R*. Accessed: 2024-11-19. Oct. 2022. URL: https://bookdown.org/mpfoley1973/survival/.

[2] Somnath Datta Glen A Satten. "The Kaplan-Meier Estimator as an Inverse-Probability-of-Censoring Weighted Average". In: *National Library of Medicine* 55.3 (Jan. 2012). Accessed: 2024-11-19, pp. 207–210. DOI: 10.1234/jar.v42i3.5678. URL: https://doi.org/110.1198/000313001317098185.

[3] Lu Tian. *Kaplan-Meier (KM) Estimator*. Accessed: 2024-11-19. Stanford University. Jan. 2014. URL: https://web.stanford.edu/~lutian/coursepdf/STAT331unit3.pdf.