The Kaplan-Meier (Product Limit) Estimate[1]

STA312 Fall 2023

[1] See last slide for copyright information.

The Kaplan-Meier Estimate

Reference: Chapter 3 in *Applied Survival Analysis Using R*

- Objective: To estimate the survival function without making any assumptions about the distribution of survival time.
- If there were no censoring, it would be easy.
- Use the empirical distribution function: the proportion of observations less than or equal to t.

$$\widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} I\{t_i \leq t\}$$

- Then let $\widehat{S}_n(t) = 1 - \widehat{F}_n(t)$

Discrete Time

Maybe time is always discrete in practice

- Consider times $t_0 = 0, t_1, t_2, \ldots,$ maybe minutes or days.
- Let $q_j =$ the probability of failing at time t_j, given survival to time t_{j-1}.
- This is the *idea* behind the hazard function.
- $p_j = 1 - q_j =$ the probability of surviving past time t_j, given survival past time t_{j-1}.

$$\begin{aligned} p_j &= P(T > t_j | T > t_{j-1}) \\ &= \frac{P(T > t_j, T > t_{j-1})}{P(T > t_{j-1})} \\ &= \frac{P(T > t_j)}{P(T > t_{j-1})} \\ &= \frac{S(t_j)}{S(t_{j-1})} \end{aligned}$$

$p_j = \frac{S(t_j)}{S(t_{j-1})}$

Probability of surviving past time t_j, given survival past time t_{j-1}

With $S(t_0) = S(0) = 1$,

- $p_1 = \frac{S(t_1)}{S(t_0)} = \frac{S(t_1)}{1} = S(t_1)$
- $p_2 = \frac{S(t_2)}{S(t_1)}$
- $p_3 = \frac{S(t_3)}{S(t_2)}$
- Continuing ...
- $p_k = \frac{S(t_k)}{S(t_{k-1})}$

Then,

$$
\begin{aligned}
& \quad\; p_1 \qquad p_2 \qquad p_3 \quad \cdots \quad p_k \\
&= S(t_1) \frac{S(t_2)}{S(t_1)} \frac{S(t_3)}{S(t_2)} \cdots \frac{S(t_k)}{S(t_{k-1})} \\
&= S(t_k)
\end{aligned}
$$

$S(t_k) = \prod_{j=1}^{k} p_j$

Estimate $S(t_k)$ by estimating the p_j.

- Let d_j be the number of deaths at time t_j.
- Let n_j be the number of individuals at risk before time t_j.
- Anyone censored before time t_j is no longer at risk.
- Estimated probability of failure at time t_j is $\widehat{q}_j = \frac{d_j}{n_j}$.

$$
\begin{aligned}
\widehat{p}_j &= 1 - \widehat{q}_j = \frac{n_j - d_j}{n_j} \\
\widehat{S}(t_k) &= \prod_{j=1}^{k} \widehat{p}_j \\
\widehat{S}(t) &= \prod_{t_j \le t} \widehat{p}_j
\end{aligned}
$$

Working toward a standard error for $\widehat{S}(t) = \prod_{t_j \le t} \widehat{p}_j$

Large-sample Distribution Theory

- $\widehat{p}_j = 1 - \frac{d_j}{n_j} = \frac{n_j - d_j}{n_j}$ is a sample proportion – a sample mean.
- It is the proportion of individuals eligible at risk for failure at time t, who did not fail.
- Mean of independent Bernoullis (conditionally on n_j).
- $E(\widehat{p}_j) = p_j$, $Var(\widehat{p}_j) = \frac{p_j(1-p_j)}{n_j}$
- $\widehat{p}_j \dot{\sim} N(p_j, \frac{p_j(1-p_j)}{n_j})$ by the Central Limit Theorem.
- This is for large n_j.

Recall

Theorem based on the delta method of Cramér

Let $\boldsymbol{\theta} \in \mathbb{R}^k$. Under the conditions for which $\widehat{\boldsymbol{\theta}}_n$ is asymptotically $N_k(\boldsymbol{\theta}, \mathbf{V}_n)$ with $\mathbf{V}_n = \frac{1}{n}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})^{-1}$, let the function $g : \mathbb{R}^k \to \mathbb{R}$ be such that the elements of $\dot{\mathrm{g}}(\boldsymbol{\theta}) = \left(\frac{\partial g}{\partial \theta_1}, \ldots, \frac{\partial g}{\partial \theta_k}\right)$ are continuous in a neighbourhood of the true parameter vector $\boldsymbol{\theta}$. Then

$$g(\widehat{\boldsymbol{\theta}}) \dot{\sim} N\left(g(\boldsymbol{\theta}), \dot{\mathrm{g}}(\boldsymbol{\theta})\mathbf{V}_n\, \dot{\mathrm{g}}(\boldsymbol{\theta})^\top\right).$$

Note that the asymptotic variance $\dot{\mathrm{g}}(\boldsymbol{\theta})\mathbf{V}_n\, \dot{\mathrm{g}}(\boldsymbol{\theta})^\top$ is a matrix product: $(1 \times k)$ times $(k \times k)$ times $(k \times 1)$.

The standard error of $g(\widehat{\boldsymbol{\theta}})$ is $\sqrt{\dot{\mathrm{g}}(\widehat{\boldsymbol{\theta}})\widehat{\mathbf{V}}_n\, \dot{\mathrm{g}}(\widehat{\boldsymbol{\theta}})^\top}$.

Specializing the delta method to the case of a single parameter

Yielding the univariate delta method

Let $\boldsymbol{\theta} \in \mathbb{R}$. Under the conditions for which $\widehat{\theta}_n$ is asymptotically $N(\theta, v_n)$ with $v_n = \frac{1}{n} I(\theta)$, let the function $g(x)$ have a continuous derivative in a neighbourhood of the true parameter θ. Then

$$g(\widehat{\theta}) \dot{\sim} N\left(g(\theta), g'(\theta)^2 \, v_n\right).$$

The standard error of $g(\widehat{\theta})$ is $\sqrt{g'(\widehat{\theta})^2 \, \widehat{v}_n}$, or $\left|g'(\widehat{\theta})\right| \sqrt{\widehat{v}_n}$

Large-sample Distribution Theory Continued

$\widehat{S}(t) = \prod_{t_j \leq t} \widehat{p}_j$ with $\widehat{p}_j = \frac{n_j - d_j}{n_j} \stackrel{.}{\sim} N\left(p_j, \frac{p_j(1-p_j)}{n_j}\right)$

- Sums are easier to work with than products.
- $\log \widehat{S}(t) = \sum_{t_j \leq t} \log \widehat{p}_j$
- Using the one-variable delta method, $\log \widehat{p}_j \stackrel{.}{\sim} N(\log p_j, \frac{1-p_j}{n_j p_j})$
- Sum of normals is normal (asymptotically, too).
- $E(\sum_{t_j \leq t} \log \widehat{p}_j) \approx \sum_{t_j \leq t} \log p_j = \log \prod_{t_j \leq t} p_j = \log S(t)$

$$
\begin{aligned}
Var\left(\sum_{t_j \leq t} \log \widehat{p}_j\right) &\approx \sum_{t_j \leq t} Var(\log \widehat{p}_j) \\
&= \sum_{t_j \leq t} \frac{1 - p_j}{n_j p_j}
\end{aligned}
$$

Asymptotic Distribution of $\log \widehat{S}(t) = \sum_{t_j \le t} \log \widehat{p}_j$

$$\log \widehat{S}(t) \,\dot{\sim}\, N\left(\log S(t), \sum_{t_j \le t} \frac{1-p_j}{n_j p_j}\right)$$

- This is a stepping stone to the distribution of $\widehat{S}(t)$.
- Use the univariate delta method again.
- Univariate delta method says that if $T_n \,\dot{\sim}\, N(\theta, v_n)$ then $g(T_n) \,\dot{\sim}\, N\left(g(\theta), v_n[g'(\theta)]^2\right)$.
- Here, $T_n = \log \widehat{S}_n(t)$, $\theta = \log S(t)$ and $g(x) = e^x$.
- $g'(\theta) = e^\theta = e^{\log S(t)} = S(t)$. So,

$$\widehat{S}(t) \,\dot{\sim}\, N\left(S(t), S(t)^2 \sum_{t_j \le t} \frac{1-p_j}{n_j p_j}\right)$$

Standard error of $\widehat{S}(t)$

Used in the denominator of Z-tests and $\widehat{S}(t) \pm 1.96\, se$

$$\widehat{S}(t) \stackrel{\cdot}{\sim} N\left(S(t), S(t)^2 \sum_{t_j \leq t} \frac{1-p_j}{n_j p_j}\right)$$

- Of course we don't know $S(t)$ or p_j in the variance.
- So use estimates.
- Estimate $S(t)$ with $\widehat{S}(t)$, and estimate p_j with $\widehat{p}_j = \frac{n_j - d_j}{n_j}$.
- The resulting estimated asymptotic variance is $\widehat{S}(t)^2 \sum_{t_j \leq t} \left(\frac{d_j}{n_j(n_j - d_j)}\right)$
- This is expression (3.1.2) on p. 27 of the text.
- The standard error of $\widehat{S}(t)$ is $\widehat{S}(t)\sqrt{\sum_{t_j \leq t} \left(\frac{d_j}{n_j(n_j - d_j)}\right)}$.
- In R's `survival` package, the default confidence interval for the Kaplan-Meier estimate uses this standard error.

Counting Processes

The theoretical state of the art

- Distribution theory for the Kaplan Meier estimate (asymptotic normality, standard error etc.) has been presented the way it was originally developed.
- The derivation is partly sound, but it has some holes.
- More recently, viewing number of failures up to a point as a counting process (stochastic processes, STA348 and beyond) has cleaned the whole thing up.
- Results are the same, but now the proofs are rigorous.
- There was some guesswork in the development of these ideas, but the main guesses were right.

Copyright Information