BIO 244: Unit 5

Kaplan-Meier (KM) Estimator

Introduction: In this section we consider the nonparametric estimation of a survivor function $S(\cdot)$ based on n i.i.d. survival times that can be noninformatively right censored. The resulting estimator–commonly known as the Kaplan-Meier Estimator or the Product-Limit Estimator–is probably one of the most commonly-used estimators in medical/public health studies involving failure time data. The development will be largely heuristic, with formal proofs of large-sample properties deferred to later units.

Suppose that $T_1, T_2, \ldots, T_n$ are i.i.d. survival times with survivor function $S(\cdot)$, with $C_1, C_2, \ldots, C_n$ the censoring times, i.i.d. and independent of the T_i, and suppose that our observations are denoted (U_i, δ_i) for $i = 1, 2, \ldots, n$, with

$$U_i = T_i \wedge C_i, \qquad \delta_i = 1_{\{T_i \leq C_i\}}.$$

To begin, let us suppose that $F(\cdot)$ is discrete with mass points at $v_1 < v_2 < \cdots$ (where $v_1 \geq 0$), and define the discrete hazard functions

$$h_1 = P[T = v_1]$$

and

$$h_j = P[T = v_j \mid T > v_{j-1}]$$

for $j > 1$.

Note that for $t \, \epsilon \, [v_j, v_{j+1})$,

$$\begin{aligned} S(t) &\stackrel{\text{def}}{=} P(T > t) = P(T > v_j) \\ &= P(T > v_j \mid T > v_{j-1}) P(T > v_{j-1}) \\ &= P(T > v_j \mid T > v_{j-1}) P(T > v_{j-1} \mid T > v_{j-2}) P(T > v_{j-2}) \\ &\quad \vdots \\ &= (1 - h_j)(1 - h_{j-1}) \cdots (1 - h_1) = \textstyle\prod_{i=1}^{j} (1 - h_i). \end{aligned}$$

Similarly, define $f_1 = h_1$ and, for $j > 1$,

$$f_j \stackrel{\text{def}}{=} P(T = v_j) = h_j \cdot \prod_{i=1}^{j-1}(1 - h_i).$$

Now consider making an inference about F based on the likelihood function corresponding to (U_i, δ_i) for i=1,2,...,n. This is just

$$L(F) = \prod_{U_i:\delta_i=1} f(U_i) \cdot \prod_{U_i:\delta_i=0} (1 - F(U_i)).$$

Substituting the h_j, this becomes (after some algebra):

$$L(F) = \prod_j h_j^{d_j}(1 - h_j)^{Y(v_j)-d_j}, \tag{5.1}$$

where $0 \leq h_j \leq 1$ and

$$d_j = \sum_{i=1}^{n} \delta_i \cdot 1_{\{U_i=v_j\}} = \#\text{ who fail at } v_j$$

and

$$Y(v_j) = \sum_{i=1}^{n} 1_{\{U_i \geq v_j\}} = \#\text{ ``at risk'' at } v_j.$$

The maximizing solution is seen to be (for $Y(v_j) > 0$):

$$\hat{h}_j = d_j / Y(v_j),$$

so that

$$\hat{S}(t) = \begin{cases} 1 & t < v_1 \\ \prod_{i=1}^{j}(1 - \hat{h}_i) & v_j \leq t < v_{j+1}. \end{cases}$$

Notice that the expression for $\hat{h}_j$ makes sense: the probability of dying at v_j given you are alive before is estimated by $d_j/Y(v_j)$. Also the expression for $\hat{S}(t)$ makes sense: the probability of staying alive at v_j if alive before v_j is estimated by $(1 - d_j/Y(v_j))$.

Example: Suppose that v_j : $2, 4, 5, 7, 9, 11, 16, 18, 20$, and the data are as follows:

$n = 10$, (ordered) observations $= 2, 2, 3^+, 5, 5^+, 7, 9, 16, 16, 18^+$, where $^+$ means censored. Then we have:

v_j	$Y(v_j)$	d_j	$\hat{h}_j$	$\hat{S}(v_j) = \prod_{i=1}^{j}(1-\hat{h}_i) = \hat{P}(T > v_j)$	
2	10	2	2/10	.8	
4	7	0	0	.8	$(= \frac{8}{10} \times 1)$
5	7	1	1/7	.69	$(= .8 \times \frac{6}{7})$
7	5	1	1/5	.55	
9	4	1	1/4	.41	
11	3	0	0	.41	
16	3	2	2/3	.14	
18	1	0	0	.14	
20	0	0	not defined	not defined	

Note 1: — Suppose that v_g denotes the largest v_j for which $Y(v_j) > 0$ (e.g, $v_g = 18$ in the example). Then either $d_g = Y(v_g)$ or $d_g < Y(v_g)$. If $d_g = Y(v_g)$, then $\hat{h}_g = 1$, and hence $\hat{S}(t) = 0$ for $t \geq v_g$. That is, the Kaplan-Meier estimator is zero beyond time v_g. On the other hand, if $d_g < Y(v_g)$, then $\hat{S}(v_g) = \hat{P}(T > v_g) > 0$ and $\hat{S}(t)$ is not defined for larger t. Here the Kaplan-Meier estimator is an incomplete distribution–the remaining mass beyond time v_g is not defined. One way to view this is that the ML estimator of $S(\cdot)$ is not unique: any survivor function that is identical to $\hat{S}(t)$ for $t \leq v_g$ maximizes the likelihood.

Note 2: — As illustrated in the example, when calculating $\hat{S}(t)$, we only need to consider those v_j for which $d_j > 0$.

Note 3: Recall that $S(t)$ is defined as being right-continuous; that is, $S(t) = P[T > t]$. While this doesn't matter for continuous T, it does for discrete distributions.

What if we didn't know, in advance, the times at which F had mass and did not necessarily want to assume that it had to be discrete? The likelihood function is just as before; i.e.,

$$L = L(F) = \prod_{i=1}^{n} \left\{ f(u_i)^{\delta_i} (1 - F(u_i))^{1-\delta_i} \right\}.$$

However, we now need to find the maximizing solution for $F \epsilon \mathcal{F} = \left\{ \underbrace{\text{all } cdf\text{'s}}_{\nearrow} \right\}$

discrete, continuous, mixed

Kaplan and Meier argue that the maximizing solution must be a discrete distribution with mass on the observed times U_i only (see exercises). The same algebra as above leads to the same form of solution as above. Notice that this means that the Kaplan Meier estimator actually puts mass only on the observed *failure* times. That is, the Kaplan-Meier (or Product-Limit) estimator of $F(\cdot)$ is

$$\hat{S}(t) = \begin{cases} 1 & t < v_1 \\ \displaystyle\prod_{i=1}^{j} \left(1 - \frac{d_i}{Y(v_i)} \right) & \text{if } v_j \leq t < v_{j+1} \end{cases} \tag{5.2}$$

where $v_1 < v_2 < \cdots$ are distinct failure (uncensored) times. Thus, we can view $\hat{S}(\cdot)$ as a nonparametric MLE of $F(\cdot)$; this is sometimes denoted NPMLE.

One alternative (equivalent) representation of $\hat{S}(t)$ is given by:

$$\hat{S}(t) = \prod_{j\,:\,v_j \le t} \left(\frac{Y(v_j) - d_j}{Y(v_j)} \right) \qquad \text{for } t \le max(v_i), \qquad (5.3)$$

where $v_1 < v_2 < \cdots$ are the distinct observed failure times.

It is instructive to think about how the Kaplan Meier estimator places mass at the observed failure times. One way of gaining insight into this is by a construction of $\hat{S}(t)$ due to Efron (1967). This is known as the 'Redistribution of Mass' algorithm (also called redistribute to the right algorithm) (for another algorithm, see Dinse 1985).

Step 1 Arrange data in increasing order, with censored observations to the right of uncensored observations in the case of ties.

Step 2 Put mass $1/n$ at each observation.

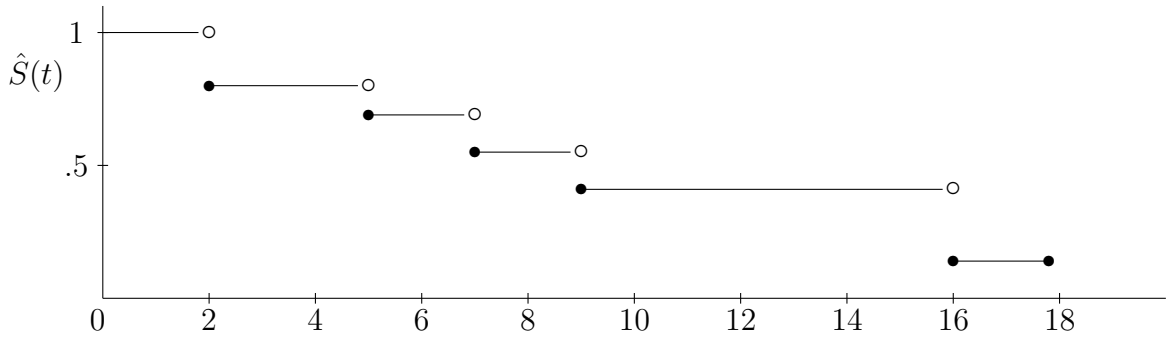
Step 3 Start from the smallest observation and move 'right'. Each time a censored observation is reached, redistribute its mass evenly to all observations to the right.

Step 4 Repeat Step 3 until all censored observations (except largest observations) have no mass. If largest (v_g) is censored, regard this mass as $> v_g$.

Let's illustrate this with the previous Example with $n = 10$.

Step 1	$\underline{2}$	$\underline{2}$	$\underline{3^+}$	$\underline{5}$	$\underline{5^+}$	$\underline{7}$	$\underline{9}$	$\underline{16}$	$\underline{16}$	$\underline{18^+}$
Step 2	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$
Step 3	$\downarrow$	$\downarrow$	$\hookrightarrow$	$\frac{1}{70}$	$\frac{1}{70}$	$\frac{1}{70}$	$\frac{1}{70}$	$\frac{1}{70}$	$\frac{1}{70}$	$\frac{1}{70}$
	$\downarrow$	$\downarrow$		$\downarrow$	$\hookrightarrow$	$\frac{1}{5}(\frac{8}{70})$	$\frac{1}{5}(\frac{8}{70})$	$\frac{1}{5}(\frac{8}{70})$	$\frac{1}{5}(\frac{8}{70})$	$\frac{1}{5}(\frac{8}{70})$
Total Mass	$\frac{2}{10}$		0	$\frac{8}{70}$	0	$\frac{24}{175}$	$\frac{24}{175}$	$\frac{48}{175}$		Assume this is somewhere > 18

$$\therefore \ \hat{S}(t) \ = \begin{cases} 1 & 0 \le t < 2 \\ .8 & 2 \le t < 5 \\ .69 & 5 \le t < 7 \\ .55 & 7 \le t < 9 \\ .41 & 9 \le t < 16 \\ .14 & 16 \le t \le 18 \\ \text{not defined} & 18 < t \end{cases}$$



Since $\hat{S}(\cdot)$ is a nonparametric estimator of $S(\cdot)$, it follows that a nonparametric estimator of $H(\cdot) = -ln(S(\cdot))$ is given by $\tilde{H}(t) = -ln(\hat{S}(t)) = -\sum_{i=1}^{j} ln(1-\hat{h}_i)$ for $v_j \le t < v_{j+1}$. However, for small x, $ln(1-x) \sim -x$ and thus this sum is approximately $\sum_{i=1}^{j} \hat{h}_i$. This suggests the alternative estimator (for $j \ge 1$):

$$\hat{H}(t) = \sum_{i=1}^{j} \hat{h}_i \qquad \text{for } v_j \le t < v_{j+1}.$$

This estimator is sometimes called the Nelson-Aalen estimator of $H(\cdot)$; we will discuss it further later in the course.

Next consider how we might approximate the distribution of $\hat{S}(t)$. One approach is to use the large-sample properties of maximum likelihood estimators, assuming that such results apply in this setting (the usual regularity conditions do not hold here since the space we are maximizing over is not a finite-dimensional parameter space). Nevertheless, let's proceed as if this is not a problem. Then from (5.1)

$$L = L(h_1, h_2, \ldots) = \prod_j h_j^{d_j} (1-h_j)^{Y(v_j)-d_j}$$

and hence

$$-\frac{\partial^2 \ln L}{\partial h_j \, \partial h_k} = 0 \qquad j \ne k$$

$$-\frac{\partial^2 \ln L}{\partial h_j^2}\bigg|_{h_j=\hat{h}_j} = Y(v_j)/(\hat{h}_j(1-\hat{h}_j))$$

$\longrightarrow$ $\hat{h}_1, \hat{h}_2, \ldots$ are approximately uncorrelated, with approximate means $h_1, h_2, \ldots$, and

$$\mathrm{Var}(\hat{h}_j) \approx \hat{h}_j(1-\hat{h}_j)/Y(v_j) = \frac{d_j(Y(v_j)-d_j)}{Y(v_j)^3}$$

Since $\hat{S}(t)$ is a product of terms of the form $1-\hat{h}_j$, it follows that $\hat{S}(t)$ is approximately unbiased in the discrete time setting as in the beginning of this unit.

For now, let us suppose that we are in the discrete time setting as in the beginning of this unit. Let's consider an estimator of the variance of the Kaplan-Meier estimator.

$$\begin{aligned}
\text{For } \; v_j \leq t < v_{j+1} \; : \; \text{Var}\left(\ln\hat{S}(t)\right) \quad &\approx \quad \sum_{i=1}^{j} \text{Var}\left(\ln(1-\hat{h}_i)\right) \\
&\overset{\delta-\text{method}}{=} \quad \sum_{i=1}^{j} \text{Var}\left(\hat{h}_i\right) \cdot \frac{1}{(1-\hat{h}_i)^2} \\
&= \quad \sum_{i=1}^{j} \frac{d_i}{Y(v_i)(Y(v_i)-d_i)}.
\end{aligned}$$

Using the δ-method again, we get what is commonly called "Greenwood's Formula"

$$\begin{aligned}
\text{Var}\left(\hat{S}(t)\right) \; &\approx \; \text{Var}\left(\ln\hat{S}(t)\right)\left(e^{\ln\hat{S}(t)}\right)^2 \\
&= \; \hat{S}(t)^2\text{Var}\left(\ln\hat{S}(t)\right) \\
&\approx \; \hat{S}(t)^2 \sum_{i=1}^{j} \frac{d_i}{Y(v_i)(Y(v_i)-d_i)} \quad (v_j \leq t < v_{j+1}).
\end{aligned}$$

One use of Greenwood's formula is to get an approximate confidence interval (e.g., a 95% CI) for $S(t)$. One obvious choice is $\hat{S}(t) \pm 1.96\sqrt{Var(\hat{S}(t))}$. However, this could give limits that are greater than 1 or less than 0. One alternative is to note that $ln(-ln\hat{S}(t))$ can take values in $(-\infty, \infty)$. Thus, using the delta method, we can approximate the variance of $ln(-ln\hat{S}(t))$ from $Var(\hat{S}(t))$, resulting in

$$Var\left(ln\left(-ln\hat{S}(t)\right)\right) \approx \frac{\sum_{i=1}^{j} \frac{d_i}{Y(v_i)(Y(v_i)-d_i)}}{\left(ln\hat{S}(t)\right)^2}.$$

Given an approximate 95% CI for $ln(-lnS(t))$ we can re-express to get the corresponding CI for $S(t)$ (see Exercises).

Breslow and Crowley (1974) show that as $n \to \infty$,

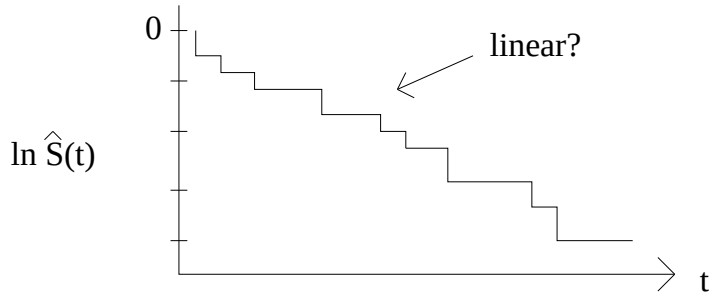
$$\sqrt{n} \quad \left(\hat{S}(\cdot) - S(\cdot)\right) \xrightarrow{w} \text{zero mean Gaussian process.}$$

Their proof is quite long and complex. We will later see that this results follows quite easily by representing the Kaplan-Meier estimator (properly transformed) as a Martingale process.

As noted earlier, the Kaplan-Meier estimator is used extensively to describe failure distributions that arise in public health and medicine. It also can be used to assess the goodness-of-fit (GOF) of a parametric assumption about F. We illustrate this with 2 examples.

Exponential Distribution: Suppose we want to check whether $T \sim Exp(\lambda)$. If it were, then $H(t) = \lambda t$, and $\ln(S(t)) = -H(t) = -\lambda \cdot t$. Hence we can plot $\ln(\hat{S}(t))$ vs t (where $\hat{S}$ is the Kaplan-Meier estimator) and visually check whether it appears linear; if so, data support the exponential assumption.

If plot appears linear, $\ln S(t) \approx -ct$
or $T \approx Exp(c)$



Weibull Distribution: Now suppose we want to check if $T \sim W(\lambda, p)$, which would imply that

$$H(t) = \lambda^p t^p \Rightarrow lnS(t) = -\lambda^p t^p$$

$$\Rightarrow ln(-lnS(t)) = p \cdot ln\lambda + plnt.$$

Thus, we can check this assumption by plotting $ln(-ln\hat{S}(t))$ vs $\ln(t)$ and checking for linearity.

Extensions to other parametric forms follow in the same way. The basic idea is to transform the c.d.f., survivor function, or integrated hazard function in a way that has a simple visual form (e.g., linearity), then replacing the $S(\cdot)$, $F(\cdot)$, or $H(\cdot)$ with its nonparametric estimator, and then checking the visual form.

SAS Commands for Kaplan Meier Estimators

Consider the dataset AZT-ddI.dat. For dataset AZTddI, with "Tad" as the time variable, "ad" as the censoring variable (ad=0 indicates censoring, ad=1 indicates event), and "rx" as the grouping variable, the following code generates Kaplan Meier curves per group:

```
proc lifetest data=AZTddI plots=(s) outsurv=aidssurv;
   time Tad *ad(0));
   strata rx;
run;
```

Outsurv=aidssurv generates a dataset aidssurv with the estimated values of the survivor function (Kaplan Meier estimator) for each group time.

For more details, see SAS help.

STATA Commands for Kaplan Meier Estimators

As with other survival analyses, one begins with the **stset** command to define the variables that represent the observed portion of survival (U) and the censoring indicator (δ): stset Tad, failure(ad) in the case of the dataset with the AZT-ddI data, where U=Tad and δ=ad.

The main STATA command is .sts. Specifically:

.sts list This gives a table with the output of the Kaplan-Meier analysis. Variations include stratified analysis by another variable; e.g. .sts list,by(gender)

.sts graph This produces a plot of the Kaplan-Meier estimator. As above, there are options such as **.sts graph,by(gender)**

sts gen varname=s This creates a new variable 'varname' whose value for subject i is $\hat{S}(u_i)$; that is, the value of the Kaplan-Meier estimator at the observed value U_i for this subject. As above, there are numerous options.

For more details, use the .help sts command in STATA.

Exercises

1. Find the K-M estimator for the following data ($n = 21$):

   $6, 6, 6, 6^+, 7, 9^+, 10, 10^+, 11^+, 13, 16, 17^+, 19^+, 20^+, 22, 23, 25^+, 32^+, 32^+, 34^+, 35^+$

   Also find an approximately 95% CI for $S(t)$ when $t = 21$.

2. Derive an algebraic expression for an approximate 95% CI for $S(t)$ that will always give an interval in $[0, 1]$. Make clear why this is an interval within $[0, 1]$!

3. Show that the K-M estimator reduces to the *ecdf* when there are no censored observations.

4. Suppose you had a random sample, say (U_i, δ_i), $i = 1, \ldots, n$, of censored survival data from a homogeneous population. Describe a graphical goodness-of-fit procedure to assess whether the hazard function for the underlying survival distribution is linear in t; that is, whether the hazard is of the form $h(t) = \alpha + \beta t$ for some unknown α, β.

   Apply this to the data above.

5. A 'natural' estimator of $H(t)$ is given by

$$\tilde{H}(t) = \sum_{j:} d_j / Y(v_j)$$

   From this, we could use the relationship $S(t) = \bar{e}^{H(t)}$ to get the estimator

$$\tilde{S}(t) = \bar{e}^{\tilde{H}(t)}.$$

   Show that this is approximately the same as the K-M estimator.

($\tilde{H}(\cdot)$ is often called the Nelson-Aalen estimator; we will consider it again later).

6. Consider the Kaplan-Meier estimator of $S(t) = P(T > t)$ based on n observations. Let $t_0 > 0$ be fixed. Prove that when the smallest censored observation exceeds t_0, the Kaplan-Meier estimator of $S(t_0)$ equals $(n - n_0)/n$, where n_0 is the number of failures prior to t_0.

7. Consider the following data set, where "+" denotes a censored value: 11.9, 6.3, 1.4+, 2.9, 7.7+, 12.4+, 4.4, 9.2+. Compute the Kaplan-Meier estimator of the survival function $S(t) = P(T > t)$. Be sure to define the estimator for all t at which it is defined.

8. Suppose we have (possibly right censored) survival data for each of two groups. Describe a graphical test for the assumption that the hazards for the 2 groups are proportional.

9. Show that $\hat{h}_j$ on page 2 is equal to $d_j/Y(v_j)$. If you use (5.1), derive (5.1), too.

10. Show that
$$-\frac{\partial^2 \ln L}{\partial h_j^2}\bigg|_{h_j=\hat{h}_j} = Y(v_j)/(\hat{h}_j(1-\hat{h}_j))$$
on page 7 holds.

11. Consider the situation on page 4, where we did not know in advance the times at which F had mass, and where we wanted to maximize the likelihood. Suppose that some potential maximizer has some mass δ in between two consecutive observations $U_{(i)}$ and $U_{(i+1)}$. What happens with the likelihood if you move this mass to $U_{(i)}$? And to $U_{(i+1)}$? Argue why this implies that the Kaplan-Meier estimator only puts mass on the U_i. In addition, argue why the Kaplan-Meier estimator does not put mass where $d_i = 0$.

12. Suppose that you plan to conduct a randomized clinical trial that compares 2 treatments, say A and B, for the prevention of recurrence of colon cancer. The outcome (response) variable is time from randomization to recurrence. It is anticipated that the enrollment (randomization)

of patients will be completed 3 years after the start of the trial, and all patients will be followed for the outcome variable from the time they are randomized until 2 years after the last patient is enrolled. Thus, depending when they enroll, each patient will be followed for 2-5 years, and her/his time until recurrence will be right censored if recurrence does not occur while he/she is followed. Patients begin their treatment immediately upon randomization. It is generally believed that if colon cancer does not recur within 3 years after beginning treatment, then it is unlikely to recur thereafter. For this reason, you are interested in testing the null hypothesis $H_0: \ S_A(3) = S_B(3)$, that the survivor functions for the 2 treatments are equal at 3 years.

Note that the choice of this null hypothesis, as opposed to something like the equality of the hazard function of the 2 groups between t=0 and t=3 years, is implicitly stressing the importance of not recurring by 3 years. The reason for this is that, under the premise of the problem, not having a recurrence by year 3 amounts to "cure". Thus, the null hypothesis implies that we don't care about differences between the hazard functions of the groups before year 3 unless these translate into a difference between the survivor functions at year 3.

(a) How would you graphically display the overall survival experiences of the patients in each treatment group? If colon cancer would not recur after 3 years if it hasn't done so by 3 years, what would you expect to see in your graph?

(b) Describe how you would test H_0. Please be specific by clearly defining a test statistic and its approximate null distribution.

References

Gregg E. Dinse (1985). An Alternative to Efron's Redistribution-of-Mass Construction of the Kaplan-Meier Estimator. *The American Statistician* Vol. 39, No. 4, Part 1, pp. 299-300.