Kaplan-Meier Survival Curve Estimation

Gary D. Knott, Ph.D.

Civilized Software Inc.
12109 Heritage Park Circle
Silver Spring, MD 20906 USA
Tel: (301) 962-3711
Email: csi@civilized.com
URL: www.civilized.com

Suppose we have survival data for a group of n similar patients, with censoring present, so that the data consists of pairs of values (y_1, e_1), (y_2, e_2), $\ldots$, (y_n, e_n) where each e_i is either 0 or 1. When $e_i = 1$, y_i is the time until death of patient i, counting from the study starting point, and when $e_i = 0$, y_i is the censoring time for patient i, indicating that patient i was lost to follow-up with an unknown fate after y_i time-units from the study starting point. The value e_i is called the result-code for patient i. Note survival time can be, in fact, the time until a "response" of some kind occurs; thus survival modeling has more general application than may be apparent.

Suppose the patients in the group have survival times, X_1, X_2, $\ldots . X_n$, which are independent identically-distributed random variables, a realization of which is, except for censoring, given by our data. Let $F(t)$ be the common distribution function of X_1, $\ldots$, X_n. We wish to estimate the survival function $S(t) = P(X_i > t) = 1 - F(t)$.

Associated with each X_i, we postulate a censoring-time random variable, C_i. The random variables C_1, $\ldots$, C_n are assumed to be independent and identically-distributed, with the distribution function $G(t) = P(C_i \leq t)$. For any realization, $(\tilde{X}_i, \tilde{C}_i)$ where $\tilde{X}_i$ is a sample of X_i and $\tilde{C}_i$ is a sample of C_i, if $\tilde{X}_i \leq \tilde{C}_i$, we have $y_i = \tilde{X}_i$, and $e_i = 1$, while when $\tilde{X}_i > \tilde{C}_i$, we have $y_i = \tilde{C}_I$ and $e_i = 0$. Thus, the value y_i is a sample of $\min(X_i, C_i)$.

If we assume a specific formula for $G(t)$, then we can estimate $S(t)$ by $(1 - H(t))/(1 - G(t))$ where $H(t)$ is the empirical cumulative distribution

function of the data-values $y_1, \ldots, y_n$.

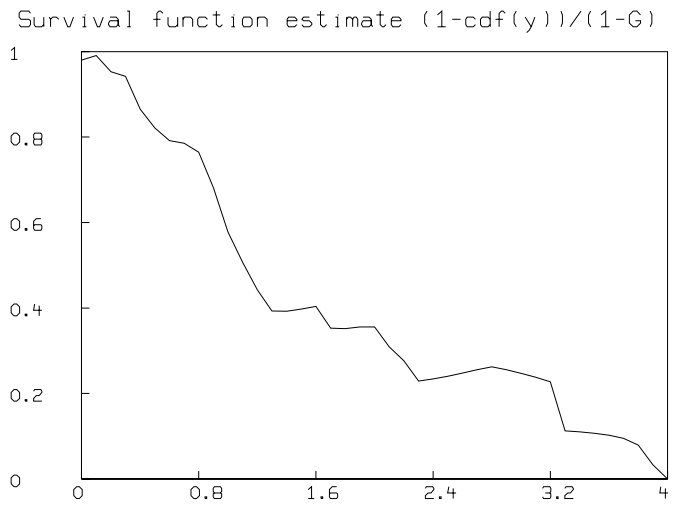
For example, suppose the censoring-time distribution is uniform on the interval $[0, 4]$; then, given the column vector Y listed below, where $Y_i = y_i$, we can estimate the survival function S in MLAB with the function SE given below. The result-code vector, E, is listed below together with Y for later reference.

```
* TYPE Y&'E

MATRIX :
1.635686 0
.3113416 0
.4963642 1
1.131409 0
2.221448 0
1.668752 1
2.785097 0
.3379156 0
.4495545 1
3.020983 1
.3994093 0
.9359379 1
.9697097 0
.6193121 0
.1234482 0
.8931070 1
.2233199 1
.8303535 0
3.926499 0
2.012626 1
.3871494 1
.9381460 1
1.210837 1
1.704919 0
1.006140 1
1.998524 1
.3763392 0
1.184550 0
.7832847 1
```

```
.9971150 0
.0452861 0
2.145687 1
.8396369 1
1.319498 0
3.288359 1
.7552886 1
1.017174 0
.8238844 1
.5893285 1
.2916714 0
3.266290 0
2.215458 1
.5546782 0
.1355474 1
.1247086 0
.9045237 1
.4080161 1
1.188316 1
1.243582 1
1.043640 1

* L = 4
* H = CDF(Y)
* FUNCTION SE(T)=IF T>=L THEN 0 ELSE (1-LOOKUP(h,t))/(1-T/L)
* DRAW POINTS(SE,0:L:.1)
* top title "Survival function estimate (1-cdf(y))/(1-G)"
* VIEW
```

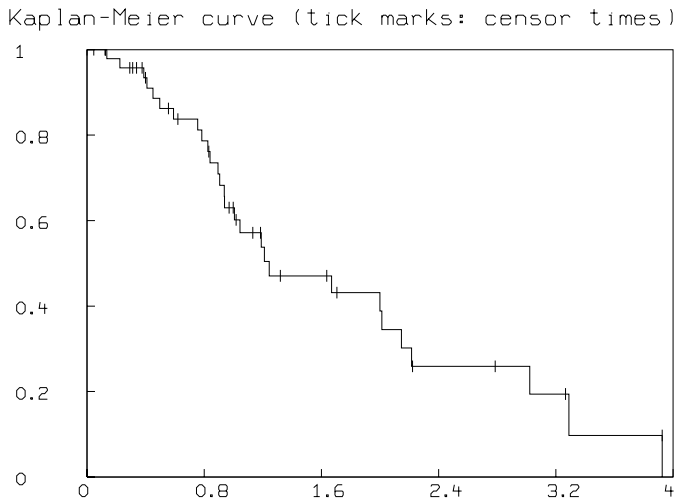


One drawback to this estimator is the fact that it is not monotonically decreasing, as is seen above. Alternatively, we can avoid assuming a specific censoring-time distribution by using the asymptotically-unbiased Kaplan-Meier product-limit estimator function, $\hat{S}$ as the estimator function for the survival function S. This can be computed in MLAB using the `KMSURV` function and graphed using the `STEPGRAPH` function.

For example, given the data, Y, listed above, together with the associated result-code vector, E, we can produce a graph of $\tilde{S}$ as shown below. The tic-marks show the points where censored observations occur.

```
* DELETE W
* D = Y&'E
* D = SORT(SORT(D,2,-1),1)
* H1 = KMSURV(D)
* H = STEPGRAPH(H1 COL 1:2)
* R = (0 &' 1) & H & (H[NROWS(H),1] &' 0)
* DRAW R, COLOR RED
* Y1 = COMPRESS(D,2,1) COL 1
* FCT F(X) = LOOKUP(H,X)
* H2 = POINTS(F,Y1)
* DRAW H2 LINE NONE, PT VBAR, PTSIZE .015, COLOR GREEN
```

```
* WINDOW 0 TO 4, 0 TO 1
* TOP TITLE "Kaplan-Meier curve (tick marks: censor times)"
* VIEW
```

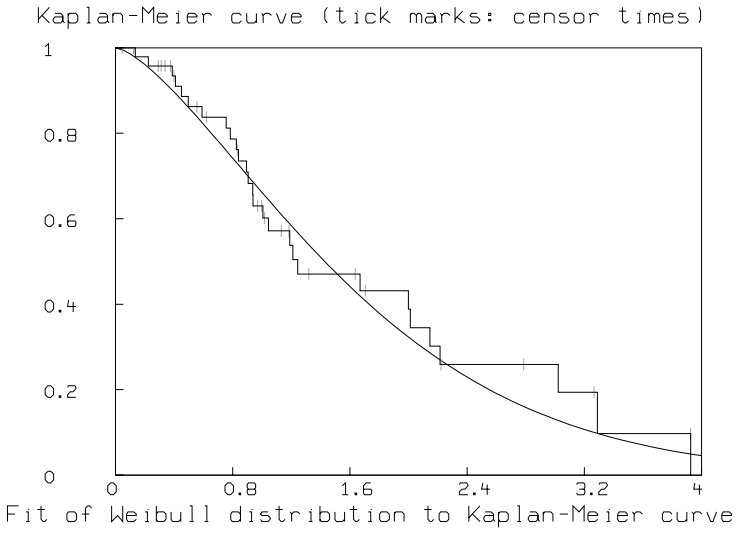


We may postulate a specific form for the distribution of the X_i random variables. For example, if the survival time distribution function is a Weibull distribution with group-specific parameters, a and b, then the survival curve is given by $SW(t) = exp(-(t/a)^b)$.

Now, we may estimate the parameters a and b using MLAB to fit the model function $SW(t)$ to a matrix of points lying on the Kaplan-Meier estimated survival curve. This is demonstrated for the matrix $H1$ of points on the estimated survival curve obtained above. Greenwood's variance approximation for $\hat{S}$ computed by `KMSURV()` in $H1$ *col 3* is used to compute appropriate weights for the curve-fit.

```
* FUNCTION SW(T)=EXP(-((IF T=0 THEN .000001 ELSE T)/A)^B)
* A = 1;B = 1; H1[1,3] = 1
* FIT(A,B), SW TO H1 COL 1:2 WITH WEIGHT 1/(H1 COL 3)
final parameter values
   value            error             dependency    parameter
```

```
    1.837800012    0.05250877741     0.3849266596   A
    1.451240067    0.05755968774     0.3849266596   B
4 iterations
CONVERGED
best weighted sum of squares = 8.691990e+00
weighted root mean square error = 5.673848e-01
weighted deviation fraction = 1.869069e-02
R squared = 9.756895e-01
* DRAW POINTS(SW,0:4:.05)
* bottom title "Fit of Weibull distribution to Kaplan-Meier curve"
* view
```



Often, we wish to compare the survival curves of several dissimilar groups of patients, to determine, for example, which of several distinct treatments is superior. This can be done by comparing the estimated survival curves directly with an appropriate statistical test (such as the Mantel-Haensel test, `MHT` provided by MLAB. We may also want to fit explanatory models to the survival curves, as above, but with embedded expressions which depend upon the various auxillary parameters, such as age, sex, or treatment. These fits then numerically characterize how the auxillary parameters determine survival time.

It may be of interest to compute the expected survival time of a subject who has already survived k time-units. This is just $M(k) := \int_k^\infty tq(t|k)\,dt$, where $q(t|k)$ is the conditional density function, $d[P(X_1 \geq t|X_1 \geq k)]/dt$. But given the survial time distribution function F and the corresponding density function $f(t) = dF(t)/dt$, we can compute $q(t|k) =$ if $t < k$ then 0 else $f(t)/a$, where

$$a = \int_k^\infty f(s)\,ds = 1 - F(k) = S(k).$$

Note $q(t|0) = 0$ and $q(t|t)$ is just the hazard function $f(t)/S(t)$.

The expected additional survival time function $M(k)$ can be easily computed and graphed in MLAB. One use for the function $M(k)$ is to estimate lifetimes for subjects with censored survival times. By thus "completing" a data set, we obtain uncensored data which is amenable to a variety of otherwise inapplicable statistical procedures.

One interesting use of survival curve modeling is as follows. Suppose we have a survival function $s(t)$, possibly obtained by curve-fitting observed survival data. Let $s(t)$ be the fraction of machines (or components, or people, or "items") which survives at least to time t. Suppose that we wish to replace these machines on a regular schedule so as to maintain the constant level of a_0 machines in service. Note that each newly-introduced replacement batch of machines follows the same survival behavior as the original machines. We wish to compute the replacement schedule function. Note the replacement function can be used for budget projection purposes.

First we shall look at the discrete formulation of the problem. Set $s_0 = 1$ and in general let s_i be the fraction of machines which survive for at least i weeks. Let a_0 be the initial number of machines at time 0, and let a_i denote the number of machines placed in service at week i. The number of machines operating starting at week k is $a_0 s_k + a_1 s_{k-1} + \cdots + a_k s_0$. Since we wish to maintain the constant level of a_0 machines at each week, we equate a_0 and $a_0 s_k + a_1 s_{k-1} + \cdots + a_k s_0$. From this we get the following recursion equation for a_k, $0 \leq k$.

$$a_k = (a_0 - \sum_{j=1}^{k-1} a_j s_{k-j})/s_0$$

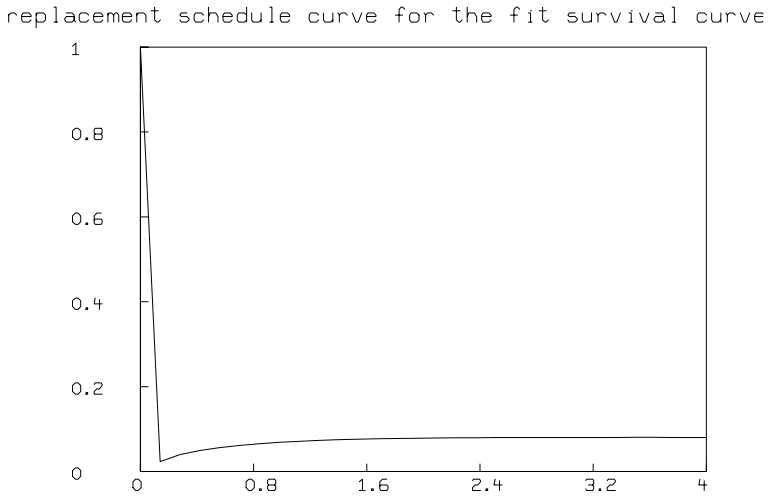
with a_0 and s_0, s_1, ..., s_k given.

This represents the solution vector of the following lower-triangular Toeplitz system.

$$\begin{bmatrix} s_0 & 0 & 0 & 0 \\ s_1 & s_0 & 0 & 0 \\ \vdots & s_1 & s_0 & \vdots \\ & \vdots & & \\ s_k & s_{k-1} & \dots & s_0 \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_k \end{bmatrix} = \begin{bmatrix} a_0 \\ a_0 \\ \vdots \\ a_0 \end{bmatrix}.$$

The MLAB function `DECONV` may be used to compute the solution to this linear system.

For example, here is the replacement schedule function computed by using `DECONV` based on the fit Weibull survival curve for the data above.



In the continuous formulation we have the given survival function $s(t)$ and we wish to compute the replacement function $a(t)$ with $a(0) = a_0$ given. The value $a(t)$ is the amount of machinery to be put into service at time t to maintain the constant level $a(0)$.

Here we get the convolution equation $a * s = a_0$ (using zero-extension). Using the Fourier convolution theorem, we get $a = ((a_0)^\wedge / s^\wedge)^\vee$. Since

the Fourier transform costs $O(k \log k)$ time for k points when k is highly composite, this deconvolution method is also a method for solving the lower-triangular system given above (when k is highly composite).