Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

A Model for Informative Censoring

William A. Link

To cite this article: William A. Link (1989) A Model for Informative Censoring, Journal of the American Statistical Association, 84:407, 749-752, DOI: 10.1080/01621459.1989.10478829

To link to this article: https://doi.org/10.1080/01621459.1989.10478829

A Model for Informative Censoring

WILLIAM A. LINK*

In the usual model for censored survival analysis, observations are of the form $X = \min(T, C)$, where T and C are nonnegative random variables representing *lifetime* and *censoring time*, respectively. Under the usual assumption of independence of T and C, the Kaplan–Meier estimator (KME) is the appropriate estimator of $S(t) = \Pr(T > t)$. The KME can lead to substantial overestimates of survival probabilities if the event of censoring indicates an unfavorable prognosis for future survival. We consider an alternative model in which censoring only occurs in a subpopulation defined by the frailty distribution. A self-consistent estimator of the survival function appropriate to the model is obtained.

KEY WORDS: Frailty; Kaplan–Meier estimator; Self-consistency.

1. INTRODUCTION

Suppose that $T_1, T_2, \ldots, T_n$ is a random sample of *lifetimes* (nonnegative continuous random variables) with common survival function $S(t) = \Pr(T > t)$. We consider the problem of estimating $S(\cdot)$ when the T's are not directly observable; rather, one is able to observe $(X_1, \delta_1), (X_2, \delta_2), \ldots, (X_n, \delta_n)$, where $X_i \leq T_i$ and δ_i is a binary random variable equaling 1 if $X_i = T_i$ and 0 otherwise.

The problem of estimating a survival function in the presence of random right censoring has been extensively studied. The majority of research has centered on the independent censoring model, in which $C_1, C_2, \ldots, C_n$ are *censoring times*, independent of $T_1, T_2, \ldots, T_n$, and $X_i = \min(T_i, C_i)$. Under this model, the Kaplan–Meier estimator (KME) (Kaplan and Meier 1958) is the appropriate estimator of $S(\cdot)$.

A great deal is known about the KME. Földes and Rejtö (1981) established strong uniform consistency of the KME, and Gill (1983) provided weak convergence results on the entire positive half-line. Furthermore, it is well known that the KME is the generalized maximum likelihood estimator of the survival function (e.g., see Miller 1981, pp. 57–59). Bivariate versions of the KME were considered by Campbell (1981) and Korwar and Dahiya (1982).

It is not difficult to envision situations in which the assumption of independent censoring is inappropriate. If the only observations available are the pairs (X, δ), however, the independence assumption is completely untestable. It has been shown by Cox (1959) and Tsiatis (1975) that "there always exist independent censoring models consistent with any probability distribution for the observable pair (X, δ)" (Lagakos 1979, p. 152). The consequence of this is that, if it is believed that the independence assumption is unwarranted, an equally untestable assumption about the joint distribution of (T, C) must be made.

As unpalatable as this may be, it is a practical necessity. Lagakos (1979) mentioned the following three situations in which the independence assumption is of questionable validity:

1. a clinical trial in which some patients remove themselves from study for reasons possibly related to therapy and thereby censor their survival time under test conditions;
2. a clinical trial in which those patients experiencing a specific critical event such as metastatic spread of disease are, by design, removed from study and no longer followed for survival time;
3. a clinical trial or animal experiment in which failure times from causes of secondary interest are recorded as censored observations of the failure times from the causes of primary interest. (p. 151)

In each of these cases, censoring indicates an unfavorable prognosis for future survival. The KME, which does not take this into account, will tend to overestimate the true survival probabilities. It is apparent, therefore, that when censoring carries an unfavorable prognosis for future survival, reasonable estimators should be bounded above by the KME and below by the empirical survival function of the observed random variable X, which we refer to as $\hat{H}(\cdot)$.

Williams and Lagakos (1977) and Lagakos and Williams (1978) defined a "cone-class" of censoring models that are indexed by a parameter $\theta \in [0, 1]$. When $\theta = 0$, censoring immediately precedes death; $\hat{H}$ is the appropriate estimator of $S(\cdot)$. When $\theta = 1$, the KME is the appropriate estimator. Their procedure for estimating $S(\cdot)$ under the cone-class model assumption involves specifying that $S(\cdot)$ is a member of some parametric family of distributions, the parameters of which, along with various model parameters, are estimated by maximum likelihood.

Another procedure for obtaining alternative estimators to the KME was proposed by Robertson and Uppuluri (1984). Their procedure is based on a modification of the well-known "redistribute to the right" algorithm due to Efron (1967). There do not appear to be easily constructible models justifying most redistribution schemes.

There would appear to be a need for some simple models and corresponding survival function estimators, applicable when censoring carries an unfavorable prognosis for future survival. In this article a model in which censoring can occur only in a "high-risk" (low-risk) subpopulation is proposed. This model suggests a modification of Efron's self-consistency algorithm that leads to a modified Kaplan–Meier estimator (MKME).

2. THE MODEL

The model for prognostic censoring considered here postulates heterogeneity in survival probabilities as de-

* William A. Link is Mathematical Statistician, Patuxent Wildlife Research Center, U.S. Fish and Wildlife Service, Laurel, MD 20708. This research was conducted while Link was a doctoral candidate at the University of Massachusetts at Amherst. The author thanks Ramesh Korwar (his doctoral advisor), the referee, and the associate editor, whose helpful comments improved the quality of the manuscript.

scribed by the "frailty model" of Vaupel, Manton, and Stallard (1979). Associated with each lifetime T is a random variable Z, called the frailty. The frailty model specifies that "the death rate or hazard at age t for a person with frailty z is assumed to be of the form $\mu(t; z) = z\mu(t)$, where $\mu(t)$ is independent of z and describes the age effect" (Hougaard 1984, p. 75). Thus conditional on frailty $Z = z$, the survival function $\Pr(T > t \mid Z = z)$ is given by

$$S(t \mid Z = z) = \exp\left\{-\int_0^t z\mu(s)\, ds\right\}. \qquad (2.1)$$

The objective is to estimate the population survival function $S(t) = E\{S(t \mid Z)\}$. The data set to be used in estimation consists of pairs (X, δ), where $X \le T$ and δ is a binary random variable equaling 1 if $X = T$ and 0 otherwise.

The independent censoring model describes the relationship between T and X by introducing a censoring time C, independent of T, and defining X as the minimum of T and C. The alternative proposed here is to assume that censoring is only possible for individuals with high (or low) frailty values. That is, let C be a *potential* censoring time: a censored observation is recorded iff $T \le C$ and Z is large (small). Letting A denote the set of values of Z for which censoring is possible, the relationship between X and T is given by

$$X = (1 - \gamma_A)T + \gamma_A \min\{T, C\},$$

where γ_A is the indicator function for the event $\{Z \in A\}$.

From (2.1) it is seen that $S(t \mid Z = z)$ is decreasing in z, so individuals with high frailties tend to have smaller lifetimes. It follows that for $A = \{z \mid z \ge a\}$, censoring will be heavier on small observations and lighter on large observations than under the independent censoring model. Under these conditions, use of the KME leads to overestimates of $S(t)$. For $A = \{z \mid z \le a\}$, the opposite effect occurs.

The model can be used to assess the robustness of the independent censoring model. In Sections 3 and 4, this is illustrated with examples in which the frailty is taken to be a unit exponential random variable. The unit exponential frailty is attractive because for any survival function $S(t)$ there exists an age effect function $\mu(t)$ that combined with a unit exponential frailty distribution yields the desired population survival function [Take $M(t)$, the integrated age effect, to be $(1 - S(t))/S(t)$.] Note that the unit exponential frailty implies that the distribution of frailty among survivors at time t is exponential with parameter $1/S(t)$, so $E(Z \mid T > t) = S(t)$.

3. SURVIVAL FUNCTION ESTIMATOR

Letting $0 = x_0 < x_1 < x_2 < \cdots < x_n$ represent the ordered times of observation and $\delta_{(1)}, \delta_{(2)}, \ldots, \delta_{(n)}$ represent the corresponding values of δ, the KME is the unique limit (as $K \to \infty$) of the sequence of functions obtained by

$$\tilde{S}^{(K+1)}(t) = \frac{1}{n}\left\{\sum_{i=1}^{n} I(x_i > t) + \sum_{i=1}^{n} (1 - \delta_{(i)})\frac{\tilde{S}^{(K)}(t)}{\tilde{S}^{(K)}(x_i)}\right\}.$$

Consequently, the KME satisfies

$$\hat{S}(t) = \frac{1}{n}\left\{\sum_{i=1}^{n} I(x_i > t) + \sum_{i=1}^{n} (1 - \delta_{(i)})\frac{\tilde{S}(t)}{\tilde{S}(x_i)}\right\},$$

which is to say that the estimated probability of survival beyond time t is the percentage of observations (censored or uncensored) beyond time t plus the estimated percentage that would have survived beyond time t but were censored before t. The KME is said to be "self-consistent" because, in the independent censoring model, for $x_i < t$,

$$\Pr(T > t \mid X = x_i, \delta = 0) = S(t)/S(x_i).$$

Under the model discussed in Section 2,

$$\Pr(T > t \mid X = x_i, \delta = 0) = \frac{S(t \mid Z \in A)}{S(x_i \mid Z \in A)},$$

suggesting that the self-consistency algorithm be replaced by

$$\tilde{S}^{(K+1)}(t) = \frac{1}{n}\left\{\sum_{i=1}^{n} I(x_i > t) + \sum_{i=1}^{n} (1 - \delta_{(i)})\left(\frac{\tilde{S}^{(K)}(t \mid Z \in A)}{\tilde{S}^{(K)}(x_i \mid Z \in A)}\right)\right\}, \qquad (3.1)$$

where $\tilde{S}^{(K)}(t \mid Z \in A)$ is the estimate of $S(t \mid Z \in A)$ based on $\tilde{S}^{(K)}(t)$. The limit as $k \to \infty$ of this sequence of survival function estimators will be referred to as the modified Kaplan–Meier estimator (MKME).

Assuming a unit exponential frailty, $S(t \mid Z \in A)$ is given by

$$S(t)\exp\left(-a\frac{1 - S(t)}{S(t)}\right),$$

for $A = \{z \mid z \ge a\}$ $(a > 0)$, and by

$$S(t)\left(\frac{1 - \exp(-a/S(t))}{1 - e^{-a}}\right),$$

for $A = \{z \mid z \le a\}$ $(a > 0)$. Substitution of these values in the algorithm described by (3.1) yields the MKME for high (low) exponential frailty censoring. Note that for $t > x$, $\Pr(T > t \mid X = x, \delta = 0)$ is less than $S(t)/S(x)$ in the case of high exponential frailty censoring (indicating that the KME will overestimate survival probabilities) and greater than $S(t)/S(x)$ in the case of low exponential frailty censoring (indicating that the KME will underestimate survival probabilities).

4. SIMULATION RESULTS

Samples of 4,000 pairs (U_T, γ_A) were generated by letting $U_T = E/(E + Z)$ and $\gamma_A = I(Z \in A)$, where E and Z are independent unit exponential variates and $A \subseteq \mathbf{R}^+$. Conditional on Z, the hazard function for U_T is $\mu(t; z) = z\mu(t)$, $t \in (0, 1)$, where the age effect is given by $\mu(t) = (1 - t)^{-2}$. Unconditionally, U_T is uniformly distributed on $(0, 1)$, so the U_T's can be thought of as the quantiles of a random sample from an arbitrary continuous survival function $S(\cdot)$. Observations of the form (U_X, δ) were then

Table 1. Survival Rates From KME (MKME) Under High Exponential Frailty Censoring (n = 4,000)

	Proportion subject to censoring								
	.10	.20	.30	.40	.50	.60	.70	.80	.90
True	Estimated								
.900	.901	.902	.903	.896	.904	.907	.902	.906	.903
	(.892)	(.902)	(.901)	(.897)	(.900)	(.900)	(.891)	(.901)	(.900)
.800	.795	.819	.805	.805	.804	.815	.809	.809	.806
	(.796)	(.800)	(.799)	(.785)	(.802)	(.790)	(.781)	(.808)	(.810)
.700	.694	.722	.716	.718	.712	.720	.723	.713	.711
	(.696)	(.693)	(.699)	(.683)	(.707)	(.682)	(.671)	(.711)	(.707)
.600	.597	.628	.620	.622	.626	.634	.630	.626	.613
	(.591)	(.596)	(.596)	(.594)	(.616)	(.570)	(.575)	(.617)	(.609)
.500	.498	.533	.527	.526	.532	.554	.546	.535	.523
	(.495)	(.494)	(.496)	(.501)	(.511)	(.476)	(.484)	(.512)	(.516)
.400	.402	.435	.440	.434	.441	.469	.463	.451	.446
	(.397)	(.398)	(.395)	(.390)	(.413)	(.389)	(.385)	(.406)	(.399)
.300	.299	.326	.338	.328	.354	.372	.374	.377	.361
	(.302)	(.295)	(.295)	(.289)	(.306)	(.301)	(.286)	(.307)	(.302)
.200	.194	.212	.230	.219	.254	.268	.286	.296	.281
	(.202)	(.200)	(.205)	(.193)	(.197)	(.201)	(.188)	(.209)	(.207)
.100	.102	.109	.112	.114	.141	.142	.161	.206	.199
	(.103)	(.094)	(.103)	(.093)	(.092)	(.102)	(.092)	(.102)	(.119)
	Proportion censored								
	.021	.053	.086	.130	.166	.222	.265	.330	.413

obtained, where

$$U_X = (1 - \gamma_A)U_T + \gamma_A \min\{U_T, U_C\}$$

and

$$\delta = 1 - \gamma_A I(U_T > U_C),$$

and the potential censoring variable U_C, uniformly distributed on (0, 1), was generated independently of E and Z. Sets A of the form $A = (a, \infty)$ and $[0, a)$ were used (for high and low exponential frailty censoring, respectively), with a chosen so that $\Pr(Z \in A) = .1(.1).9$. Results for these cases are summarized in Tables 1 and 2, respectively. As anticipated, the KME tends to overestimate survival probabilities in the former case and to underestimate survival probabilities in the latter case.

For both sets of simulations the MKME, obtained as the limit as $k \to \infty$ of the sequence given by (3.1), was computed. These results are also summarized in Tables 1 and 2.

5. DISCUSSION

The model considered in this article offers an alternative to the usual independent censoring model. The event of censoring is related to the frailty of the individual. Simulations involving an exponential frailty distribution show

Table 2. Survival Rates From KME (MKME) Under Low Exponential Frailty Censoring (n = 4,000)

	Proportion subject to censoring								
	.10	.20	.30	.40	.50	.60	.70	.80	.90
True	Estimated								
.90	.901	.904	.898	.901	.905	.894	.900	.902	.892
	(.904)	(.897)	(.900)	(.892)	(.896)	(.897)	(.900)	(.905)	(.895)
.80	.806	.804	.800	.792	.799	.791	.801	.808	.784
	(.807)	(.798)	(.801)	(.796)	(.793)	(.803)	(.792)	(.810)	(.797)
.70	.705	.709	.694	.687	.693	.685	.688	.713	.682
	(.701)	(.701)	(.704)	(.698)	(.689)	(.706)	(.698)	(.707)	(.695)
.60	.604	.599	.574	.574	.577	.570	.582	.613	.582
	(.595)	(.594)	(.595)	(.586)	(.589)	(.605)	(.596)	(.600)	(.599)
.50	.502	.488	.462	.454	.473	.462	.476	.501	.485
	(.499)	(.496)	(.499)	(.486)	(.490)	(.502)	(.493)	(.492)	(.504)
.40	.390	.369	.341	.335	.359	.341	.365	.389	.383
	(.399)	(.393)	(.405)	(.381)	(.390)	(.398)	(.398)	(.383)	(.401)
.30	.273	.253	.229	.213	.241	.226	.257	.290	.285
	(.292)	(.292)	(.297)	(.285)	(.297)	(.292)	(.305)	(.281)	(.301)
.20	.164	.140	.117	.115	.138	.127	.168	.183	.185
	(.196)	(.194)	(.202)	(.187)	(.182)	(.188)	(.186)	(.171)	(.203)
.10	.056	.041	.037	.044	.066	.066	.090	.104	.074
	(.101)	(.095)	(.088)	(.082)	(.078)	(.095)	(.099)	(.071)	(.118)
	Proportion censored								
	.802	.163	.224	.273	.343	.373	.413	.462	.478

that the KME can produce substantial errors in estimation of survival probabilities.

As pointed out by Hougaard (1984), the frailty model is overparameterized unless, perhaps, the frailty is common to several individuals. In the present setting, the value of the frailty model may lie in producing a class of survival function estimates so that the potential effect of incorrectly assuming the independent censoring model can be assessed.

[*Received December 1987. Revised February 1989.*]

REFERENCES

Campbell, G. (1981), "Nonparametric Bivariate Estimation With Randomly Censored Data," *Biometrika,* 68, 417–422.

Cox, D. R. (1959), "The Analysis of the Exponentially Distributed Lifetimes With Two Types of Failure," *Journal of the Royal Statistical Society,* Ser. B, 59, 411–421.

Efron, B. (1967), "The Two Sample Problem With Censored Data," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 4), Berkeley: University of California Press, pp. 831–853.

Földes, A., and Rejtö, L. (1981), "Strong Uniform Consistency of the Product-Limit Estimator Under Variable Censoring," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete,* 58, 95–108.

Gill, R. (1983), "Large Sample Behavior of the Product-Limit Estimator on the Whole Line," *The Annals of Statistics,* 11, 49–58.

Hougaard, P. (1984), "Life Table Methods for Heterogeneous Populations: Distributions Describing the Heterogeneity," *Biometrika,* 71, 75–83.

Kaplan, E. L., and Meier, P. (1958), "Nonparametric Estimation From Incomplete Observations," *Journal of the American Statistical Association,* 53, 457–481.

Korwar, R. M., and Dahiya, R. C. (1982), "Estimation of a Bivariate Distribution Function From Incomplete Observations," *Communications in Statistics, Part A—Theory and Methods,* 12, 887–897.

Lagakos, S. W. (1979), "General Right Censoring and Its Impact on the Analysis of Survival Data," *Biometrics,* 35, 139–156.

Lagakos, S. W., and Williams, J. S. (1978), "Models for Censored Survival Analysis: A Cone Class of Variable-Sum Models," *Biometrika,* 65, 181–189.

Miller, R. G. (1981), *Survival Analysis,* New York: John Wiley.

Robertson, J. B., and Uppuluri, V. R. R. (1984), "A Generalized Kaplan–Meier Estimator," *The Annals of Statistics,* 12, 366–371.

Tsiatis, A. (1975), "A Nonidentifiability Aspect of the Problem of Competing Risks," *Proceedings of the National Academy of Science,* 72, 20–22.

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979), "The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality," *Demography,* 16, 439–454.

Williams, J. S., and Lagakos, S. W. (1977), "Models for Censored Survival Analysis: Constant-Sum and Variable-Sum Models," *Biometrika,* 64, 215–224.