A SMOOTHED VERSION

OF THE KAPLAN-MEIER ESTIMATOR

Agnieszka Rossa

Dept. of Stat. Methods, University of Lódź, Poland
Rewolucji 1905, 41, Lódź
e-mail: agrossa@krysia.uni.lodz.pl

and

Ryszard Zieliński

Inst. Math. Polish Acad. Sc.
P.O.Box 137 Warszawa, Poland
e-mail: rziel@impan.gov.pl

ABSTRACT

The celebrated Kaplan-Meter estimator (KME) suffers from a disadvantage: it may happen that estimated probabilities of survival for two different times t_1 and t_2 are equal each to other while t_1 and t_2 differ substantially. We propose a smoothinq of KME in such a way that the resulting estimator is a strictly decreasing function of time. The smoothed KME appears to be more accurate than the original one.

1. INTRODUCTION

The celebrated Kaplan-Meter estimator (KME) suffers from a disadvantage: it may happen that estimated probabilities of survival for two different times t_1 and t_2 are equal each to other while t_1 and t_2 differ substantially. It is a consequence of the fact that KME, like typical empirical distiribution function, is piecewise constant. The disadvantage has been recognized since long ago and some smoothed versions have been, explicitly or implicitly, presented in the literature. Typical approach is to choose a smooth and strictly decreasing parametric representation for

the survival probablity and to estimate that from observations at hand. For example exponential and Weibull models has been used in Greenhouse and Silliman[1], Gompertz model in Gieser *et al.*[2]), logistic, log-logistic and Weibull in Hauck *et al.*[3]. Biganzoli *et al.*[4] presented a smoothed estimate of the discrete hazard function through artificial neural network (ANN) developed as Partial Logistic regression models with ANN (PLANN). A smooth prediction through a parametric transformation of the time axis is discussed in Byers *et al.*[5]. An interesting nonparametric smoothing for survival distribution with strictly decreasing probability distribution function one can find in Xu and Prorok[6]. The literature is abundant; to not overload our note with quotations we confine ourselves to the most recent results presented in *Statistics in Medice*.

Our proposal for smoothing KME is to approximate a slightly modified version of KME **locally** by a suitable Weibull survival function. In practice it means that we fit the Weibull curve to two adjoining jump points of the original KME. The resulting estimator is a strictly decreasing function of time. It appears to be more accurate than the original KME.

2. MODEL AND ESTIMATION

We assume a nonparametrical model: the survival probablity function is any continuous and strictly decreasing function $F(t)$ for $t \geq 0$ with $F(0) = 1$ and $\lim_{t\to\infty} F(t) = 0$. Typical representatives are exponential, Weibull, gamma, generalized gamma, lognormal, Gompertz, Pareto, log-logistic, and exponential-power distribtuions, to mention the most popular among them (see e.g. Kalbfleisch and Prentice[7] , Klein *et al.*[8]). Every survival probability function may be locally approximated with a prescribed level of accuracy by a Weibull $W(t;\lambda,\alpha)$ survival probability function of the form $W(t;\lambda,\alpha) = exp\{-\lambda t^{\alpha}\}$. For that reason we construct our estimator, to be denoted by $S_2(t)$ (a reason for the subcript will become clear later), as follows.

Denote by $t_1, t_2, \ldots, t_N$ the jump points of KME, by $P_1, P_2, \ldots, P_N$ the values of KME at those points, and by $\bar{P}_1, \bar{P}_2, \ldots, \bar{P}_N$ the arithmetic means of KME in close left-hand and right-hand vinicities of a given point t (by the very definition,

at the point t_i KME jumps down from the level P_{i-1} to the level P_i (we define $t_0 = 0$ and $P_0 = 1$). Hence we define $\bar{P}_i = (P_{i-1} + P_i)/2$ for $i = 1, 2, \ldots, N-1$; for $i = N$ we define $\bar{P}_N = P_N/2$ if the last observation is censored and $\bar{P}_N = P_N$ otherwise. We shall illustrate our considerations using the well known data on the effect of 6-mercaptopurine on the duration of steroid-induced remission in acute leukemia taken from Freireich *at al.*[9] (see also Marubini and Valsecchi[10]). The "survival times" of 21 clinical patients were

$$6, 6, 6, 6^*, 7, 9^*, 10, 10^*, 11^*, 13, 16, 17^*, 19^*, 20^*, 22, 23, 25^*, 32^*, 32^*, 34^*, 35^* \quad (1)$$

where * denotes a censored observation. Kaplan-Meier estimator for that data is presented in Fig. 1 and in the following table:

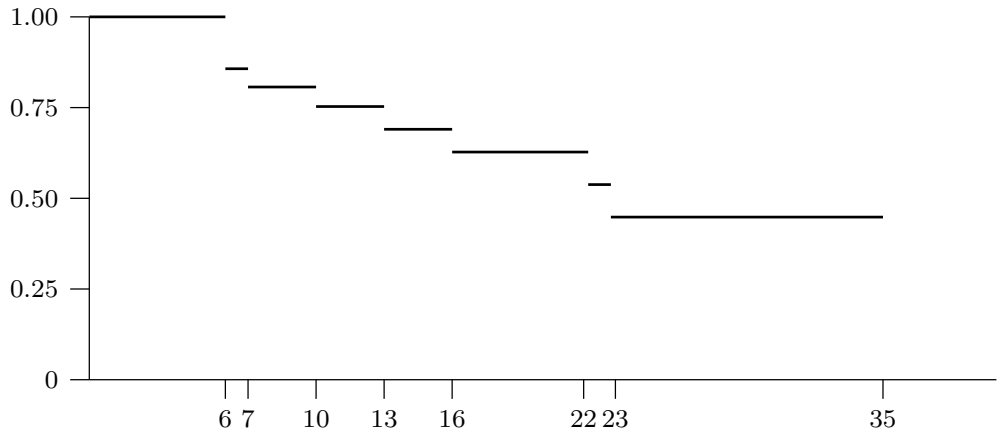

Fig.1. Kaplan-Meier estimator for data (3)

Tab.1. KME and modified KME for data (1)

i	1	2	3	4	5	6	7	8
t_i	6	7	10	13	16	22	23	35
P_i	.857	.807	.753	.690	.627	.538	.448	.448
$\bar{P}_i$	.928	.832	.780	.722	.659	.583	.493	.448

To estimate the survival probability for a given t we define our estimator $S_2(t)$ as follows.

If $t = t_i$ for some i, then $S_2(t) = \bar{P}_i$.

If $0 < t \leq t_N$ and $t_i < t < t_{i+1}$ then we choose a Weibull survival probability function which pass through the points $(t_i, \bar{P}_i)$ and $(t_{i+1}, \bar{P}_{i+1})$ and then as the value of our estimator $S_2(t)$ we take the value of the fitted Weibull survival probability function at that point t. It amounts to finding values of λ and α, say $\hat{\lambda}$ and $\hat{\alpha}$, such that

$$W(t_i; \hat{\lambda}, \hat{\alpha}) = \bar{P}_i \quad and \quad W(t_{i+1}; \hat{\lambda}, \hat{\alpha}) = \bar{P}_{i+1} \tag{2}$$

Then $S_2(t) = W(t; \hat{\lambda}, \hat{\alpha})$. Solving (1) amounts to solving, with respect to Λ and α, the simple set of two linear equations

$$\begin{cases} \alpha \, \log t_i + \Lambda = \log(-\log \bar{P}_i) \\ \alpha \, \log t_{i+1} + \Lambda = \log(-\log \bar{P}_{i+1}) \end{cases} \tag{2'}$$

with $\Lambda = \log \lambda$.

If $t > t_N$ than we proceed as follows:

— if the last observed t_N is a censoring time, our estimator, like the original KME, is not defined;

— otherwise we solve (2) for $i = N - 1$ (we extrapolate the Weibull curve which is based on two largest not censored observations).

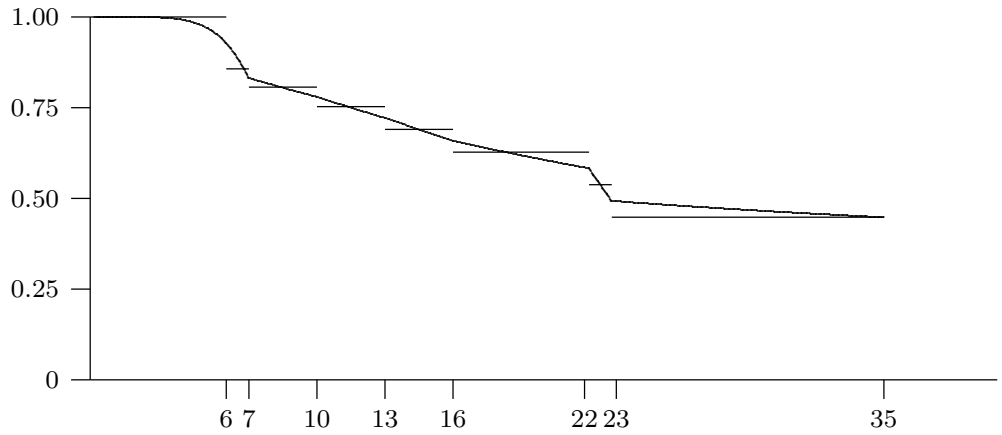

Fig.4. Kaplan-Meier and S_2 estimators for data (3)

Estimator $S_2(t)$ for data (1), as well as original KME, are presented in Fig. 2. For example, if $t = 25$ or $t = 33$ the original KME gives us the predicted survival equal to 0.448 in both cases, while our estimator gives us $S_2(25) = 0.484$ and $s_2(33) = 0.456$, respectively. Similarly, for $t = 17$ and $t = 20$ KME is equal to 0.627, $S_2(17) = 0.645$, and $S_2(20) = 0.607$. Between the two points KME is constant while $S_2(t)$ strictly decreases.

3. SIMULATION

To assess teh accuracy of the new estimator we performed a great number of computer simulations. It appeared that Mean Square Error and Mean Absolute Deviation were significantly smaller. Also Pitman's Measure of Closeness advocates for our estimator. Detailed numerical results are given in a technical report (Rossa and Zieliński[11]) which we can sent to an interested reader in a TeX-file form.

4. DISCUSSION

The proposed estimator $S_2(t)$ is based on a local fitting a Weibull survival probability to two neighbouring step points of the modified KME. One could expect that a similar estimator $S_k(t)$ based on k neighbours would perform better. It evidently gives us a better smoothing but any interval on time axis which contains $k > 2$ neighbouring points is of course larger than that for $k = 2$ which may result in a poorer local approximation of an unknown survival curve from a nonparametric family by a Weibull one. Also some practical questions arise: instead of solving (2) or (2′) one has to apply a technic of fitting two-parameter Weibull curve to $k > 2$ points, for example a version of the least square method. All these advocate for a very simple but still quite satisfactory estimator $S_2(t)$.

REFERENCES

1. Greenhouse, J.B., and Silliman, N.P. 'Applications of a mixture survival model with covariates to the analysis of depression prevention trial' SM 15, 2077-2094 (1996)

2. Gieser, P.W., Chang, M.N., Rao, P.V., Shuster, J.J., and Pullen, J. 'Modelling cure rates using the Gompertz model with covariate information' SM 17, 831-839 (1998)

3. Hauck, W.W., McKee, L.J., and Turner, B.J. 'Two-part survival models applied to administrative data for determining rate of and predictors for maternal-child transmission of HIV' SM 16, 1683-1694 (1997)

4. Biganzoli, E., Boracchi, P., Mariani, L., and Marubini, E. 'Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach', *Statistics in Medicine*, 17, 1169-1186 (1998)

5. Byers, R.H. Jr., Caldwell, M.B., Davis, S., Gwinn, M., and Lindegren, M.L. 'Projection of AIDS and HIV incidence among children born infected with HIV' SM 17, 169-181 (1998)

6. Xu, J.-L. and Prorok, P.C. 'Non-parametric estimation of the post-lead-time survival distribution of screen-detected cancer cases', *Statistics in Medicine*, **14,**, 2715–2725 (1995).

7. Kalbfleisch, J.D. and Prentice, R.L. 'The statistical analysis of failure time data'. Wiley (1980)

8. Klein, J.P., Lee, S.C. and Moeschberger, M.L. 'A partially parametric estimator of survival in the presence of randomly censored data' *Biometrics*, 46, 795–811 (1990).

9. Freireich, E.O. *et al.* 'The effect of 6-mercaptopurine on the duration of steroid-induced remission in acute leukemia: a model for evaluation of other potentially useful therapy', *Blood*, **21**, 699–716 (1963).

10. Marubini, E. and Valsecchi, M.G. 'Analysing Survival Data from Clinical Trials and Observational Studies', Wiley (1995).

11. Rossa, A. and Zieliński, R. 'Locally Weibull–Smoothed Kaplan–Meier ESTimator', *Institute of Mathematics Polish Academy of Sciences*, Preprint 599, November 1999.