Selecting A Threshold for Long Term Survival Probabilities and Kaplan-Meier Estimator

Authors: Ion Grama, Jean-Marie Tricot & Jean-François Petiot

Speaker: Patrick Thompson

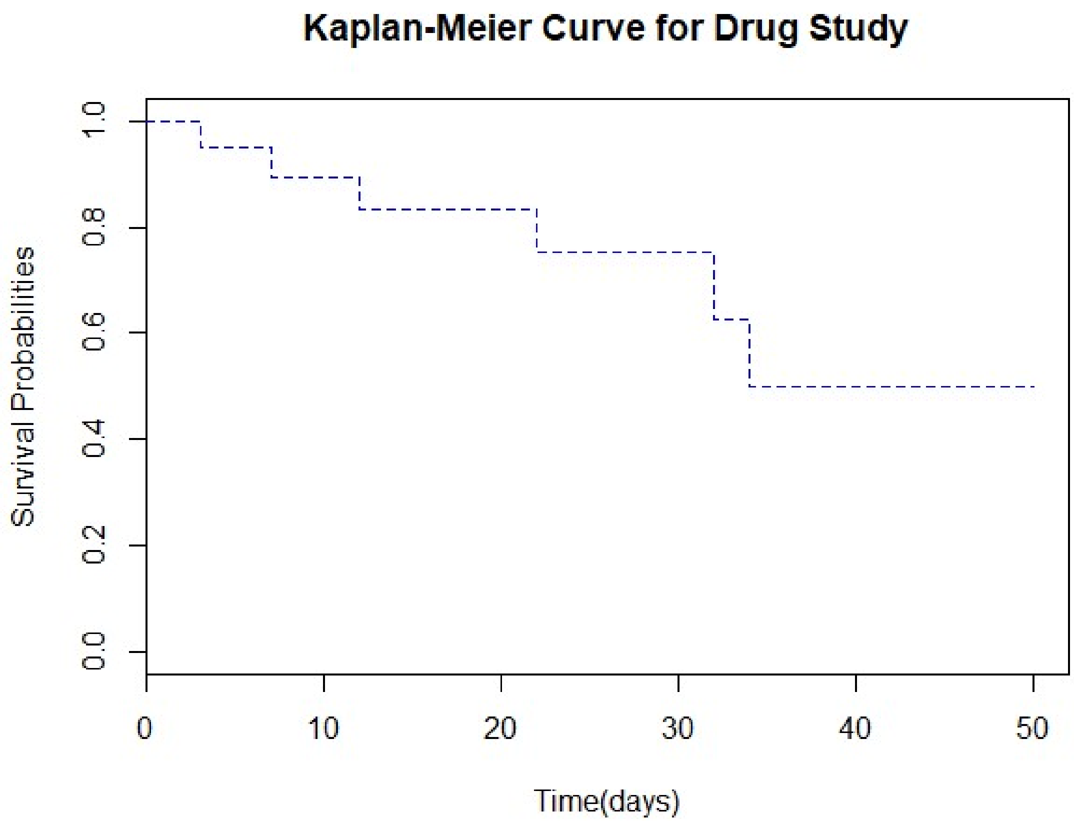
Venue: University of Texas at Dallas

Date: Thursday, December 6, 2018

Table of Contents

- Motivation for Paper
- Aim of Paper
- Model and Background Definitions and Assumptions
- Consistency of Estimator with Fixed Threshold
- Homework Problem

Motivation for Paper



Motivation for Paper

```
library(survival)

time<-c(3,5,6,7,10,12,15,18,19,20,22,25,27,29,32,34,38,42,44,50)

delta<-c(1,0,0,1,0,1,0,0,0,0,1,0,0,0,1,1,0,0,0,0)

data.surv<-survfit(Surv(time,delta)~1)
plot(data.surv,lty=2,col="blue",xlab="Time(days)",ylab = "Survival Probabilities",conf.int =
    F)

title("Kaplan-Meier Curve for Drug Study")
```

Motivation for Paper

- Suppose that the investigator decided to terminate study after r out of the n subjects died and sacrifice the remaining n – r subjects at that time
- The survival times for the n subjects are

$$t_{(1)} \le t_{(2)} \le \ldots \le t_{(r)} = t^{+}_{(r+1)} = \ldots = t^{+}_{(n)}$$

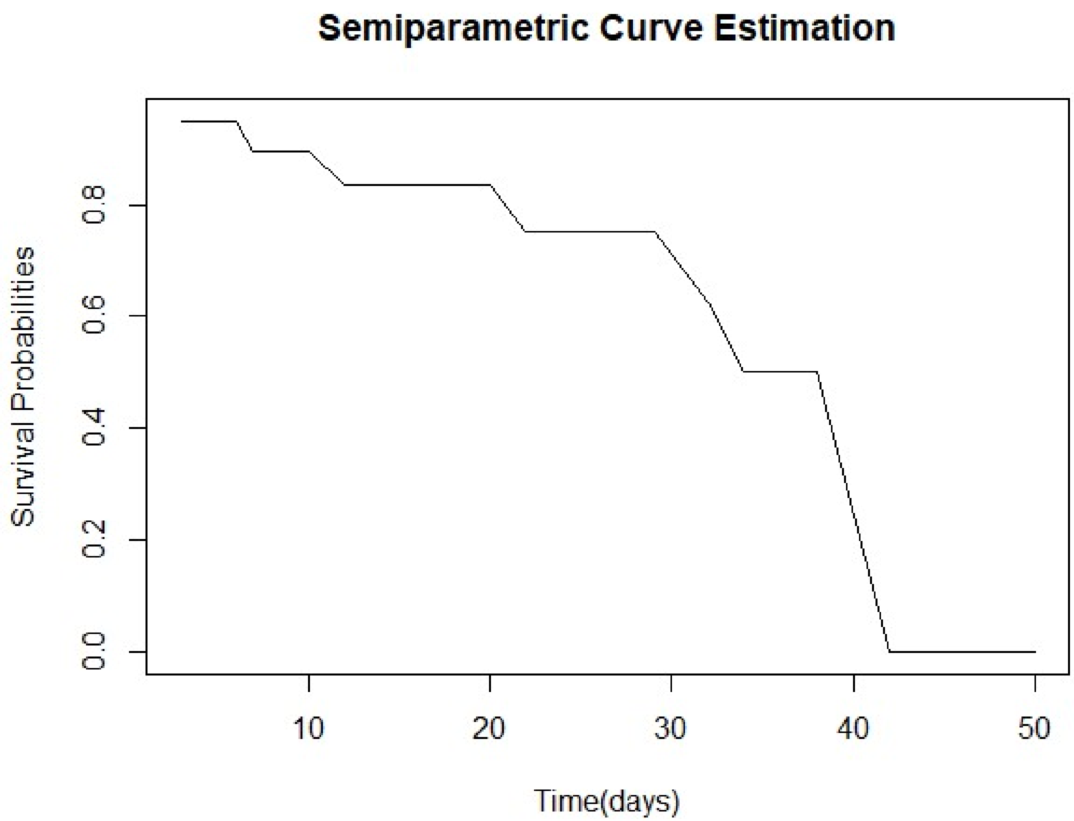
- Assuming Exp(λ) distribution, the likelihood function is $L = \frac{n!}{(n-r)!} \prod_{i=1}^{r} \lambda e^{-\lambda t_{(i)}} \left[e^{-\lambda t_{(r)}} \right]^{n-r}$

Motivation for Paper

- The MLE of λ is $\hat{\lambda} = \frac{r}{\sum_{i=1}^{r} t_{(i)} + \sum_{i=r+1}^{n} t_{(i)}^{+}}$
- Estimated mean survival time m = 1/λ is

$$\hat{\mu} = \frac{1}{\hat{\lambda}} = \frac{\sum_{i=1}^{r} t_{(i)} + \sum_{i=r+1}^{n} t_{(i)}^{+}}{r}$$

Motivation for Paper



Motivation for Paper

```
mu_hat<-(sum(time[c(1,4,6,11,15,16)])+sum(time[17:20]))/4fx <- (x > 0 & x < 34
    )*data.surv$surv + (x > 34 & x < 50)*exp(-mu_hat*x)

plot(x, fx,type = "l",xlab="Time(days)",ylab = "Survival
    Probabilities")title("Semiparametric Curve Estimation")

title("Semiparametric Curve Estimation")
```

Aim of Paper

- Kaplan-Meier (KM) Nonparametric Estimator One of Most Popular for Estimating Survival Time Distributions in Right Censoring
- KM Estimator May be Unreliable in Estimating Extreme Survival Probabilities if Censoring Rate is High
- KM Estimator is More Flexible & Preferable

  (Meier et al, 2004)

Aim of Paper

- Possible Solution is to Combine KM Estimator and a Parametric-Based Model Into One Construction
- Completely Using a Parametric Model Can Lead to a High Bias if Misspecified
- Fit the Tail of the Survival Function with a Parametric Model
- Otherwise, Continue to Use KM Estimator

Aim of Paper

- Propose a Procedure for Automatic Choice of Tail Location Based on G.O.F. Test
- Technique Will Improve the Estimation of Survival Probabilities in Mid and Long Term
- New Estimator Incorporates Threshold t Using KM Estimator for times up t & By Parametric Model, Otherwise
- Parametric Model Chosen is Exp(θ)

Important Model & Background Definitions & Assumptions

- Assume Survival & Right Censoring Times Arise From Non-negative Variables X & C
- X & C May Depend On A Categorical Covariate Z
- Let $f_F(\cdot|z)$ & $S_F(\cdot|z) = 1 - F(\cdot|z)$ be the conditional density & survival functions of X given Z = z.

Important Model & Background Definitions & Assumptions

- Corresponding conditional hazard function is

  $h_F(\cdot|z) = f_F(\cdot|z)/S_F(\cdot|z)$ given $Z = z$
- C has conditional density $f_C(\cdot|z)$, survival function $S_C(\cdot|z)$ and hazard function

  $h_C(\cdot|z) = f_C(\cdot|z)/S_C(\cdot|z)$ given $Z = z$
- Assume X & C are independent, conditionally w.r.t. Z

Important Model & Background Definitions & Assumptions

- Let observation time & failure indicator be $T = \min\{X, C\}$ and $\Delta = 1_{\{X \leq c\}}$
- Let $P_{F,Fc}\,(dx, d\delta|z)$, $x \geq x_0 \geq 0$, $\delta \in \{0, 1\}$ be the conditional distribution of the vector $\mathbf{Y} = (T, \Delta)'$ given $Z = z$
- Density of $P_{F,Fc}$ is $P_{F,Fc}\,(x, \delta|z) =$ $f_F(x|z)^{\delta}\, S_F(x|z)^{1-\delta} f_c(x|z)^{1-\delta} S_C(x|z)^{\delta}$

Consistency of Estimator

- Define quasi-log likelihood by

$$L_t(\theta \mid z) = \sum_{i=1}^{n} \log p_{F_{\theta,t},F_C}\left(T_i, A_i \mid z_i\right) 1_{\{z_i = z\}},$$

where $F_{\theta,t}\left(x \mid z\right) = \begin{cases} F(x \mid z), x_0 \le x \le t \\ 1-\left(1-F(t \mid z)\right)\exp\left(-\frac{x-t}{\theta}\right), x > t \end{cases}$

with parameters $\theta > 0$, $t \geq x_0$ and $F(\cdot|z) \in \mathscr{F}$, $z \in Z$

Consistency of Estimator

- Using the definition of $p_{F,\,Fc}$ and dropping censoring terms, partial quasi-log likelihood is

$$L_t^{part}(\theta \mid z) = \sum_{T_i \le t, z_i = z} \Delta_i \log h_{F_{\theta,t}}(T_i \mid z)$$

$$- \sum_{T_i > t, z_i = z} \Delta_i \log\theta - \sum_{T_i \le t, z_i = z} \int_{x_0}^{T_i} h_{F_{\theta,t}}(v \mid z) dv$$

$$- \sum_{T_i > t, z_i = z} \left(\int_{x_0}^{t} h_{F_{\theta,t}}(v) dv + \theta^{-1}(T_i - t) \right)$$

for fixed $z \in z$ and $t \geq x_0$

Consistency of Estimator

- Maximizing $L_t^{part}(\theta \mid z)$ in θ yields the estimator $\hat{\theta}_{z,t} = \sum_{T_i > t, z_i = z} (T_i - t) / \hat{n}_{z,t}$, where by convention 0/0 = ∞ and $\hat{n}_{z,t} = \sum_{T_i > t, z_i = z} \Delta_i$ is the number of observed survival times beyond the threshold t.
- Estimator of $S_F(x)$, $x_0 \leq x \leq t$ is obtained by standard nonparametric ML approach due to Kiefer and Wolfowiz (1956)

Consistency of Estimator

- Use the product KM estimator (with ties) defined by $\hat{S}_{KM}(x \mid z) = \prod_{T_i \le x} \left(1 - d_i(z)\right) / r_i(z)$,
  x ≥ x_0, where $r_i(z) = \sum_{j=1}^{n} 1_{\{T_j \ge T_i, z_j = z\}}$
  is the number of individuals at risk at T_i &
  $d_i(z) = \sum_{j=1}^{n} 1_{\{T_j = T_i, \Delta_j = 1, z_j = z\}}$
  is the number of individuals who died at T_i

Consistency of Estimator

- Semiparametric fixed-threshold KM estimator (SFKM) of survival function takes form

$$\hat{S}_t(x \mid z) = \begin{cases} \hat{S}_{KM}(x \mid z), x_0 \le x \le t \\ \hat{S}_{KM}(t \mid z) \exp\left(-\frac{x-t}{\hat{\theta}_{z,t}}\right), x > t \end{cases},$$

where $\exp(-(x-t)/\hat{\theta}_{z,t}) = 1$ if $\hat{\theta}_{z,t} = \infty$

Consistency of Estimator

- Denote by $n_z = \sum_{i=1}^{n} 1(z_i = z)$ the number of individuals with profile $z \in Z$
- Assume that there is a constant $\kappa \in (0, 1]$ such that for any $z \in Z$, $n_z \geq \kappa n$
- Let P be the joint distribution of the sample Y_i, $i = 1, \ldots, n$ and E be the expectation with respect to P

Consistency of Estimator

- The notation $\alpha_n = O_P(\beta_n)$ means there is a positive constant c such that

  $P(\alpha_n > c\beta_n, \beta_n < \infty) \rightarrow 0$ as $n \rightarrow \infty$, for any two sequences of positive possibly infinite variables α_n and β_n
- Consider the Kullback-Leibler divergence

  $K(\theta', \theta) = \int \log(dG_{\theta'} / dG_{\theta}) dG_{\theta'}$ between two exponential distributions with means θ' & θ

Consistency of Estimator

- By convention, $K(\infty,\theta)=\infty$
- $K(\theta',\theta)=\psi(\theta'/\theta-1)$, with $\psi(x)=x-\log(x+1), x>-1$

  and there are two constants c_1 and c_2 such that

  $(\theta'/\theta-1)^2 \le c_1 K(\theta',\theta) \le c_2 (\theta'/\theta-1)^2$,

  when $|\theta'/\theta-1|$ is small enough

Consistency of Estimator

- Theorem:

  Assume that $n_z \geq \kappa n$, $h_F(\cdot|z)$ satisfies
  $|\theta_z h_F(\theta_z x|z) - 1| \leq A\exp(-\alpha_z x)$, $x \geq x_0$, $A > 0$,
  $\theta_{max} > \theta_{min} > 0$ be constants, $\alpha_z > 0$ and
  $\theta_z \in (\theta_{min}, \theta_{max})$ and $h_C(\cdot|z)$ satisfies
  $|\theta_z h_C(\theta_z x|z) - \gamma_z| \leq M(1 + x)^{-\mu}$, $x \geq x_0$, $M > 0$,
  $\gamma_{max} > \gamma_{min} > 0$ be constants, $\mu > 1$,
  $\gamma_z \in (\gamma_{min}, \gamma_{max})$

Consistency of Estimator

- Theorem Cont'd

  Then, $K\left(\hat{\theta}_{z,t_{z,n}},\theta_z\right) = O_p\left(\left(\frac{\log n}{n}\right)^{\frac{2\alpha_z}{1+\gamma_z+2\alpha_z}}\right)$, where

  $t_{z,n} = \frac{\theta_z}{1+\gamma_z+2\alpha_z}\log n + o(\log n)$
- Assume that the survival time X is exponential, i.e. $h_F(x \mid z) = \theta_z^{-1}$ for all $x \geq x_0$ and $z \in Z$
- The assumption ensures supposition 2 in the theorem

Consistency of Estimator

- Assume suppositions 1 and 3 in the theorem
- Let there be constants θ_{min} and θ_{max} such that $0 < \theta_{min} \leq \theta_z \leq \theta_{max} < \infty$
- The theorem implies $\left|\hat{\theta}_{z,t_{z,n}} - \theta_z\right| = O_P\left(\left(n^{-1}\log n\right)^{\frac{\alpha}{1+\gamma_z+2\alpha}}\right)$ for any $\alpha > 0$
- This rate becomes arbitrarily close to the $n^{-1/2}$ rate as $\alpha \rightarrow \infty$ since $\lim_{\alpha\to\infty} \alpha/(1+\gamma_z+2\alpha) \rightarrow 1/2$

Consistency of Estimator

- Thus the estimator $\hat{\theta}_{z,t_{z,n}}$ almost recovers usual parametric rate of convergence as a becomes large whatever is $\gamma_z > 0$
- In the case when there is no censoring ($\gamma_z = 0$) after an exponential rescaling, problem can be reduced to estimation of extreme index
- If $\gamma_z \rightarrow 0$ our rate becomes close to $n^{-\frac{2\alpha_z}{1+2\alpha_z}}$ known to optimal in extreme value estimation (Dress, 1998 & Gramma & Spokoiny, 2008)

Homework Problem

- Suppose that in a laboratory experiment 10 mice are exposed to carcinogens. The experimenter decides to terminate the study after half of the mice are dead and to sacrifice the other half at that time. The survival times of the five expired mice are 4, 5, 8, 9, and 10 weeks. The survival data of the 10 mice are 4, 5, 8, 9, 10, 10+, 10+, 10+, 10+, and 10+. Assuming that the failure of these mice follows an exponential distribution, estimate the survival rate λ and mean survival time μ.

References

1. Dirk F. Moore. Applied Survival Analysis Using R. Switzerland:Springer, 2016.
2. Elisa T. Lee, John Wenyu Wang. Statistical Methods for Survival Data Analysis. 4th ed. Hoboken, NJ: John Wiley & Sons, Inc., 2013
3. Faye Anderson. Survival Analysis by Example Hands On Approach Using R. 1st ed.
4. Ion Grama, Jean-Marie Tricot, Jean-François Petiot. Long Term Survival Probabilities and Kaplan-Meier Estimator. 14 pages. 2013. <hal-00918822>