Exam. Survival Analysis 2020-2021

January 18th, 2021

The exam consists of three exercises, Exercise 1, Exercise 2 and a Practical that is a continuation of Exercise 2. They weight 20% , 50% and 30%, respectively.

Exercise 1

Consider two one-dimensional covariates X and Z that are binary and independent. Given these we assume that a survival time T has hazard function $\alpha(t; X, Z)$. Further $P(X = 1) = \pi_x$ and $P(Z = 1) = \pi_z$ with $0 < \pi_x < 1$ and $0 < \pi_z < 1$.

(a) Given that the hazard function is on Cox form

$$\alpha(t; X, Z) = \alpha_0(t) \exp(X\beta + Z\gamma)$$

then find $P(X = i, Z = j|T > t)$, the covariate distribution among survivors. Is it possible to choose β and γ to make X and Z conditionally independent given $T > t$?

(b) Given that the hazard function is

$$\alpha(t; X, Z) = \alpha_0(t) + X\beta + Z\gamma + XZ\rho$$

then find $P(X = i, Z = j|T > t)$. Is it possible to choose β, γ and ρ to make X and Z conditionally independent given $T > t$?

Exercise 2

Let T be an exponentially distributed random variable with

$$P(T > t) = e^{-\theta t},$$

where $\theta > 0$ is an unknown parameter. Let U be a random variable independent of T with

$$P(U > t) = \exp\left(-\int_0^t \gamma(s)\, ds\right),$$

where $\gamma(t)$ is an unknown hazard function. We assume that U is always observed but for T we only know whether $T \in [0, U)$. Hence we observe the pair (U, δ), where $\delta = I(T < U)$. Define the counting processes $N_j(t) = I(U \leq t, \delta = j)$, $j = 0, 1$.

(a) Show that

$$P(\delta = 1|U = t) = 1 - e^{-\theta t}$$

and use this to argue that the two counting process $N_j(t)$ has intensity $\lambda_j(t) = I(t \leq U)\alpha_j(t)$, $j = 0, 1$, where

$$\alpha_0(t) = \gamma(t)e^{-\theta t}$$
$$\alpha_1(t) = \gamma(t)\left\{1 - e^{-\theta t}\right\}.$$

Let (U_i, δ_i) be n iid random variables from the above generic setting and define $N_{ij}(t)$, for $j = 0, 1$, and $i = 1 \dots, n$. Define also, for $j = 0, 1$,

$$N_{\cdot j}(t) = \sum_{i=1}^{n} N_{ij}(t)$$

and $N(t) = N_{\cdot 0}(t) + N_{\cdot 1}(t)$. Similarly, let $M_{ij}(t) = N_{ij}(t) - \int_0^t Y_i(t)\alpha_j(s)\,ds$ where $Y_i(t) = I(t \leq U_i)$. Define also $M_{\cdot j}(t) = \sum_{i=1}^{n} M_{ij}(t)$ and $M(t) = M_{\cdot 0}(t) + M_{\cdot 1}(t)$.

(b) Argue that $M_{i0}(t)$ and M_{k1} for $1 \leq i, k \leq n$ are orthogonal (square-integrable) martingales and show that $M_{\cdot 0}(t)$ and $M_{\cdot 1}(t)$ are orthogonal.

(c) Derive the compensator for $N(t)$ and use that to argue that a natural estimator of $\Gamma(t) = \int_0^t \gamma(s)\,ds$ is

$$\hat{\Gamma}(t) = \int_0^t \frac{1}{Y(s)} dN(s),$$

where $Y(t) = \sum_{i=1}^{n} Y_i(t)$.

(d) Argue that, when we observe in $[0, t]$, that the likelihood function is given by

$$L_t = \exp\left\{-\int_0^t Y(s)\gamma(s)\,ds\right\} \prod_{i=1}^{n} \left[\gamma(U_i)\left\{1 - e^{-\theta U_i}\right\}^{\delta_i} e^{-\theta U_i(1-\delta_i)}\right]^{I(U_i \leq t)}$$

Let

$$U_t(\theta) = \frac{\partial}{\partial \theta} \log L_t$$

(e) Show that

$$U_t(\theta) = \int_0^t \frac{s e^{-\theta s}}{1 - e^{-\theta s}} dN_{\cdot 1}(s) - \int_0^t s N_{\cdot 0}(s)$$

and find its compensator. Use this to argue that $U_t(\theta)$ must be a martingale.

Let $\Gamma^*(t) = \int_0^t \gamma(s) I(Y(s) > 0)\,ds$ and define $\tilde{M}(t) = \hat{\Gamma}(t) - \Gamma^*(t)$. We assume that we observe in the interval $[0, \tau]$ with $\tau < \infty$ and we let $\hat{\theta}$ denote the solution to $U_\tau(\theta) = 0$. All time points t considered in the following are assumed to be within $[0, \tau]$.

(f) Show that $U_t(\theta)$ and $\tilde{M}(t)$ are orthogonal and calculate $\langle U_t(\theta) \rangle$.

(g) Show, under appropriate assumptions, that $n^{1/2}\tilde{M}(t)$ converges in distribution towards a Gaussian martingale as $n \to \infty$, and show that the variance function of the limiting process is $\frac{P(U \leq t)}{P(U > t)}$.

(h) Show, under appropriate assumptions, that $n^{-1/2}U_t(\theta)$ converges in distribution towards a Gaussian martingale as $n \to \infty$, and derive the variance function of the limiting process.

(i) Use the result in (h) to derive the asymptotic distribution of $n^{1/2}(\hat{\theta} - \theta)$, show that the variance, $\sigma^2(\theta)$, of the limiting distribution is

$$\sigma^2(\theta) = \left(\int_0^\tau \frac{s^2 e^{-\theta s}}{1 - e^{-\theta s}} f_U(s)\,ds\right)^{-1}$$

where $f_U(s)$ denotes the density function of the distribution of U.

A natural estimator of $\sigma^2(\theta)$ is

$$\hat{\sigma}^2(\theta) = \left(\int_0^\tau \frac{s^2 e^{-\hat{\theta} s}}{1 - e^{-\hat{\theta} s}} e^{-\hat{\Gamma}(s)} \frac{1}{Y(s)} dN(s) \right)^{-1}$$

(j) Argue that $\hat{\sigma}^2(\theta)$ is a consistent estimator of $\sigma^2(\theta)$.

Practical (Exercise 2 continued)

We shall now investigate how the above estimator and its corresponding standard error estimator performs in practice. We wish to simulate data from a situation where both T and U are exponentially distributed with mean 1, ie. we take $\theta = 1$. We let the observation interval be given by $\tau = 1$. We wish to study the performance of the two estimators where we take the sample size n equal to 400. You may use the following R-code to generate data

```
n=400
T=rexp(n,1)
U=rexp(n,1)
delta=as.numeric(T<U)
tau=1

## Data that we observe

U.obs=U*as.numeric(U<tau)+tau*as.numeric(U>=tau)
delta.obs=delta*as.numeric(U.obs<tau)+999*as.numeric(U>=tau) # value 999 corresponds to
                                                             # the unobserved values of delta.
status=as.numeric(U.obs<tau)

## In the following you can only use the observed data.
## ie: (U.obs,delta.obs,status)
```

In the below questions (a) and (b) you only need one random sample generated as suggested above.

(a) Calculate and plot $\hat{\Gamma}(t)$. In the same figure you should also plot the straight line with intercept 0 and slope 1. Comment on the plot. Help: consider whether you can use the `aalen`-function.

(b) Calculate $\hat{\theta}$ and $\hat{\sigma}(\theta) = \sqrt{\hat{\sigma}^2(\theta)}$. Keep in mind that $\hat{\theta}$ is the zero-root of $U_\tau(\theta)$ and to find this you can for instance use the function `nleqslv` in the R-package `nleqslv`. To calculate $\hat{\sigma}(\theta)$ it may be helpful to use the `aalen`-function.

In the following, `theta.tot` and `see.tot` are supposed to contain the calculated values of $\hat{\theta}$ and $\hat{\sigma}(\theta)$ based on 2000 runs.

(c) Now you should report the following results:

```
mean(theta.tot)
sd(theta.tot)
mean(see.tot)
```

What can you conclude from this?