Exam: Survival analysis

October 20, 2021

This is the exam for the Survival Analysis course 2021. You can work in groups of up to three persons, state your names clearly on the first page.
The exam consist of 2 theoretical exercises and one practical part, each excercise count for 1/3 of the points.
Exams should be send to `ts@biostat.ku.dk` before 11 November 2021, 12:00.

1 Theoretical part

1.1 Marginal Models

Consider waiting times $T_1 = W_1$ and $T_2 = T_1 + W_2$. We shall be interested in the marginal models for W_1 and W_2 given covariates X. We assume, given X and an independent random effect ρ that is Gamma-distributed with mean 1 and variance θ, that the two waiting times W_1 and W_2 are independent such that

$$P(W_1 > w_1|\rho, X) = \exp(-\Lambda_1(w_1)\rho\exp(X^T\beta_1))$$
$$P(W_2 > w_2|\rho, X) = \exp(-\Lambda_2(w_2)\rho\exp(X^T\beta_2)),$$

where $\Lambda_j(w_j) = \int_0^{w_j} \lambda_j(s)ds$ are two increasing baseline functions, and β_j are regression coefficients, for $j = 1, 2$. Note that given only X the waiting times W_1 and W_2 are dependent.
We observe the waiting times T_1, T_2 subject to independent right-censoring, such that given X, ρ we have that C is independent of T_1, T_2. Further, the conditional distribution of C given X, ρ does not depend on ρ. We thus observe $\tilde{T}_1 = \min(T_1, C)$, $\delta_1 = I(T_1 < C)$, $\tilde{T}_2 = \min(T_2, C)$, $\delta_2 = I(T_2 < C)$ and X.
We assume that we have i.i.d. observations from this generic model.

1. What is the marginal survival distribution of W_1 given X, $P(W_1 > w_1|X)$, and what is the related hazard function.

2. What is the observed hazard function of W_1 given X and ρ, i.e., what is

$$\lim_{h\to 0}\frac{1}{h}P(W_1 \in [t, t+h]|W_1 > t, C > t, X, \rho)$$

   - a. What is the observed hazard function of W_1 given only X.
   - b. Can we estimate $\Lambda_1()$ and β_1 based on the observed data.

3. What is the observed hazard of W_2 given X and ρ, i.e., what is

$$\lim_{h\to 0}\frac{1}{h}P(W_2 \in [t, t+h]|W_2 > t, C > W_1 + t, X, \rho)$$

- a. Compute $H(w_1, w_2|X) = P(W_1 \leq w_1, W_2 > w_2|X)$ using the underlying Gamma distribution that models the dependence. This function will appear in the expression for the observed hazard function.
- b. What is the observed hazard function of W_2 given only X, i.e., what is

$$\lim_{h} \frac{1}{h} P(W_2 \in [t, t+h] | W_2 > t, C > W_1 + t, X)$$

Hint: You may interchange limits, differentation and integration as you like for the computations without additional arguments.

4. This question does not rely on the particular structure set up earlier, but has a more general non-parametric point of view. Compute

$$L(w_1, w_2) = E\left(\frac{I(W_1 \leq w_1, W_2 > w_2)I(C > W_1 + w_2)}{G_c(W_1 + w_2|X)}\right)$$

for $w_1 \leq K_1$ and $w_2 \leq K_2$, with $G_c(t|X) = P(C > t|X)$ and assuming that $G_c(w|x) > \delta$ for all x and $w \leq K_1 + K_2$.

a. Assume that we know $G_c(\cdot|X)$, then suggest and estimator of $L(w_1, w_2)$ for $w_1 \leq K_1$ and $w_2 \leq K_2$ based on i.i.d. observations, and use this to estimate $P(W_1 > w_1)$ for $w_1 \leq K_1$.

b. Suggest an estimator of $P(W_2 > w_2|W_1 \leq w_1)$ for relevant (w_1, w_2).

1.2 Marginal Models for recurrent events

Let D denote a survival time (the terminal event), and let $N^*(t)$ count the number of recurrent events observed over a time-period $[0, t]$, where $t \leq \tau$. Due to the terminal event, we only observe the recurrent event processes up to $\tau \wedge D$, where $a \wedge b = \min(a, b)$. Clearly subjects will only have events when still alive. Observations may also be censored, thus only making it possible to observe the processes up to the censoring time C. Let $\delta = I(D \leq C)$, $T = D \wedge C$, and $N(t) = N^*(t \wedge T)$ be the observed number of events and define the at-risk process $Y(t) = I(T \geq t)$. Denote the counting process of the terminal event by $N^D(t)$ and denote its marginal cumulative hazard by $\Lambda^D(t)$. We make the standard assumption that the censoring time, C, is independent of D and $N^*(t)$. Note, however that D and $N^*(t)$ may be dependent.
The observations $\{N_i(t) : t \leq \tau, T_i, \delta_i\}$ are assumed to be independent replicates of $\{N(t) : t \leq \tau, T, \delta\}$ for $i = 1, ..., n$.
We consider the martingale associated to the censoring time, $M_i^C(t) = N_i^C(t) - \int_0^t Y_i(s) d\Lambda^C(s)$, with counting process $N_i^C(t) = I(T_i \leq t, \delta = 0)$ and with at-risk indicator $Y_i(t) = I(D_i > t, C_i > t)$. Where as usual $\Lambda^C(t) = \int_0^t \lambda_c(s) ds$ is the cumulative hazard for C and denoting the survival function of C by $G_c(t) = P(C > t)$ such that $P(C > \tau) > \epsilon > 0$ with ϵ some constant. We shall consider an estimator of the marginal mean $\mu(t) = E(N^*(t \wedge D))$,

$$\hat{\mu}(t) = \frac{1}{n} \sum_i \int_0^t r_i(s) I(D_i \geq s) dN_i(s)$$

with $r_i(s \wedge D_i) = I(C_i \geq s \wedge D_i)/G_c(s \wedge D_i)$ being un-censored at time t. We assume that G_c is known to simplify some calculations.

1. Show that the estimator is an unbiased estimator of $\mu(t)$, when considering a t such that $G_c(t) > \delta > 0$.

2. Show that

$$r_i(s) = 1 - \int_0^s \frac{1}{G_c(u)} dM_i^C(u),$$

by calculating the right hand side the expression.

3. Based on this show that the estimator can be written as

$$\hat{\mu}(t) = \frac{1}{n} \sum_i \left\{ \int_0^t I(D_i \geq s) dN_i^*(s) - \int_0^t H_i(s,t) \frac{1}{G_c(s)} dM_i^C \right\}$$

where $H_i(s,t) = \int_s^t I(D_i > u) dN_i^*(u)$.

4. Show that the variance of

$$\rho_i(t) = \left\{ \int_0^t I(D_i \geq s) dN_i^*(s) - \mu(t) - \int_0^t H_i(s,t) \frac{1}{G_c(s)} dM_i^C \right\}$$

is

$$E \left\{ \int_0^t I(D_i \geq s) dN_i^*(s) - \mu(t) \right\}^2 + \int_0^t E(H_i(s,t) I(D_i > s))^2 \frac{\lambda_c(s)}{G_c(s)} ds$$

Hint: one option is to work with a larger filtration (including $H_i(s,t)$ and D_i) for the censoring martingale such that the predictable variation can be computed using the standard martingale formula.

5. What is the best function that we can compute, i.e., that depends on the observed data on the form $\epsilon_i(t) = \int_0^t \alpha_i(s) \frac{1}{G_c(s)} dM_i^C$ to add to $\rho_i(t)$ to the minimize the variance of $\rho_i(t) + \epsilon_i(t)$. So give an expression for $\alpha_i(s)$, where $\alpha_i(s)$ is some function of the observed data up to time s. Such that we can use the variance calculation as in 4.
Technical Hint: $\alpha_i(s)$ is a predictable function given the observed history such that $\epsilon_i(t)$ is a martingale with respect to the observed history.

   a. What is the best function $\alpha_i(s)$ when it is not allowed to depend on the data.

   b. How can this be used to estimate the marginal mean $\mu(t)$.

   Hint: the conditional mean, $\phi(X) = E(Y|X)$ will minimize $E(Y - \phi(X))^2$ among functions ϕ.

2 Practical part

2.1 Colrectal Cancer

We consider the colorectal cancer data from the frailtypack R-package. This gives the new lessions prior to death during follow-up of metastatic colorectal cancer patients. The data consist of 150 patients from the follow-up of the FFCD 2000-05 multicenter phase III clinical trial originally including 410 patients with metastatic colorectal cancer randomized into two therapeutic strategies: combination and sequential. The dataset contains times of observed appearances of new lesions censored by a terminal event (death or right-censoring) with baseline characteristics (treatment arm, age, WHO performance status (who.ps) and previous resection).

```
library(frailtypack)
data(colorectal)
head(colorectal);
```

```
  id     time0     time1 new.lesions treatment        age who.PS
1  1 0.0000000 0.7095890           0         S 60-69 years      0
2  2 0.0000000 1.2821918           0         C   >69 years      0
3  3 0.0000000 0.5245902           1         S 60-69 years      1
4  3 0.5245902 0.9207650           1         S 60-69 years      1
5  3 0.9207650 0.9424658           0         S 60-69 years      1
6  4 0.0000000 0.6639344           1         C 60-69 years      0
  prev.resection state   gap.time
1             No     1 0.70958904
2             No     1 1.28219178
3             No     0 0.52459017
4             No     0 0.39617486
5             No     1 0.02170073
6            Yes     0 0.66393442
```

We shall be interested in the covariate effects of treatment, age, and who.ps We use follow up time as the main time-scale. Part of this exercise is to figure out how to answer the posed question, and draw relevant conclusions and also consider and state important assumptions.

1. Show if covariates are important for death using Cox modelling. Estimate the survival distribution for relevant covariate combinations.

2. Is death and the number new lesions related ?

3. Estimate the mean number of new lessions as a function of time, i.e., the marginal mean of the recurrent events ($\mu(t)$ in the previous exercise).

4. Estimate the probability of a patient having more than one new lession before dying as a function of time.

    - Investigate if covariates are important for this probability by fitting an appropriate regression model.
    - Remember to check assumptions and validate the regression model.

5. Estimate the probability of a patient having more than two new lessions before dying as a function of time.

    - Investigate if covariates are important for this probability.