UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in: STK4080 — Survival and event history analysis

Day of examination: Thursday 10 June 2021

Examination hours: 09.00 – 13.00

This problem set consists of 6 pages.

Appendices: None

Permitted aids: Approved calculator

Please make sure that your copy of the problem set is complete before you attempt to answer anything.

Problem 1 Hospital length of stay of COVID-19 patients

In the beginning of the corona pandemic, a group of Chinese medical researchers used data from confirmed COVID-19 patients at hospitals in the Sichuan Province. They collected information on demographic, epidemiological, clinical characteristics, and the length of hospital stay for these patients. The study included 538 patients who were admitted in hospitals between January 16 and March 4, 2020. 332 out of these (62%) recovered and were discharged before the end of the study, April 4, 2020. Only 3 patients died in hospital before April 4.

The data used in this Problem are *simulated* based on the reported results from the study (the full underlying data were not reported in the study). They consist of the observed time, `Time` (in days); censoring status `D` (= 0 or 1); and six binary covariates $\mathtt{x}_1, \ldots, \mathtt{x}_6$ (with values 0 and 1 as defined by Table 1 and explained below).

Here, for patients discharged at or before April 4, `Time` is the *true length* of hospital stay. These patients are given censoring status $\mathtt{D} = 1$. For patients that were alive and still in hospital on April 4, `Time` is the *observed length* of stay. These patients are given censoring status $\mathtt{D} = 0$. For patients who died during hospital stay, `Time` is the time from hospital admission to death. These patients are also given censoring status $\mathtt{D} = 0$.

In Table 1, 'time from onset' means time from onset of COVID-19 to admission at the hospital; 'hospital grade' distinguishes between admission to a provincial (rural) and non-provincial (urban) hospitals; 'density of health workers' means number of health workers per 1000 inhabitants; 'clinical grade' is degree of illness.

i	x_i	0	1
1	age (years)	< 45	≥ 45
2	gender	male	female
3	time from onset	< 5	≥ 5
4	hospital grade	non-provincial	provincial
5	density of health workers	< 5.5	≥ 5.5
6	clinical grade	mild	severe

Table 1: List of covariates for the length of stay data

(Continued on page 2.)

Below, the data for a randomly selected subset of 16 of the 538 patients is displayed .

	x1	x2	x3	x4	x5	x6	Time	D
1	0	1	0	0	1	1	3	0
2	0	1	1	0	1	0	6	1
3	0	0	1	0	1	0	10	1
4	0	0	0	1	0	1	12	0
5	0	0	0	0	0	0	12	1
6	0	1	0	0	0	0	13	1
7	0	0	1	0	0	0	14	0
8	0	0	1	0	0	1	18	1
9	0	1	0	0	1	0	21	1
10	1	1	1	0	1	0	4	0
11	1	1	0	0	1	0	8	0
12	1	0	0	0	0	0	13	0
13	1	1	0	0	0	0	18	1
14	1	1	0	0	1	0	19	1
15	1	1	0	0	0	0	22	1
16	1	0	1	0	0	0	26	0

a) Use the displayed data to calculate the Kaplan-Meier estimator separately for the two age groups < 45 and ≥ 45 (i.e., $x_1 = 0$ and 1, respectively), disregarding the other covariates.

Draw the two curves in the same figure.

What are the estimated median lengths of stay for the two groups of patients, < 45 and ≥ 45 years? What are the corresponding estimated lower and upper quartiles?

How would you compute estimated restricted mean lengths of stay for the two groups when you consider the first three weeks after admission to hospital? You need not do the full calculation. (Answers: 14.67 when $x_1 = 0$; 19.75 when $x_1 = 1$).

What would you conclude from these analyses regarding the influence of age on length of hospital stay?

The following is an edited output of a Cox-regression in R, using data for all the 538 patients and including all the covariates $x_1, \ldots, x_6$.

```
Call:
coxph(formula = Surv(Time, D == 1) ~ x1 + x2 + x3 + x4 + x5 + x6)

  n= 538, number of events= 332

        coef   se(coef)      z Pr(>|z|)
x1 -0.438713   0.120904 -3.629 0.000285
x2  0.089060   0.112596  0.791 0.428963
x3  0.004456   0.115056  0.039 0.969108
x4 -0.250050   0.126474 -1.977 0.048032
x5  0.300428   0.112477  2.671 0.007562
x6 -0.464376   0.157129 -2.955 0.003123
```

(Continued on page 3.)

b) Write down the estimated relative risk function for discharge from hospital. Which covariates have a significant effect? (Use significance level 5%).

Compute and give a practical interpretation of the estimated hazard ratios for the covariates for age and clinical grade. Derive an approximate 95% confidence interval for the hazard ratio for age.

What is the estimated relative risk of a patient of age ≥ 45 with severe clinical grade compared to a patient of age < 45 with mild clinical grade, when the other covariates are the same for the two?

The covariates for *age* and *clinical grade*, i.e., x_1 and x_6, were of particular interest in the study. In the rest of the Problem we therefore disregard the covariates $x_2, \ldots, x_5$.

c) A couple of model checks for the application of Cox regression were made for this setting:

1. It was checked whether there is an interaction between the covariates x_1 and x_6. Discuss briefly how such an interaction should be interpreted. Then explain how you would do a formal check of this interaction. No calculations are asked for here. *(No significant interaction was found in the study).*
2. It was checked whether the proportional hazards assumption is appropriate for the model with the two covariates x_1 and x_6. Explain how this could be done in the present case.

d) An alternative analysis of the data uses *Aalen's additive hazards model*. State this model when x_1 and x_6 are the only covariates.

Figure 1 shows plots from fitting the model in R.

Explain how the plots should be interpreted. What is in particular the interpretation of the Intercept plot to the left in Figure 1?

Problem 2 The Gehan-Breslow test

Consider two counting processes $N_1(t)$ and $N_2(t)$ with intensity processes of the multiplicative form

$$\lambda_h(t) = Y_h(t)\alpha_h(t); \quad h = 1, 2$$

Here, the $Y_h(t)$ are predictable processes, while the $\alpha_h(t)$ are deterministic positive functions. It is assumed that the two processes do not make simultaneous jumps.

We want to test the null hypothesis

$$H_0 : \ \alpha_1(t) = \alpha_2(t) \ \text{ for } 0 \leq t \leq t_0$$

where t_0 is the upper time limit of the study. If H_0 holds, then the common (but unknown) version of the $\alpha_h(t)$ will be called $\alpha(t)$. A subscript $\bullet$ means the sum over 1 and 2.

The test statistic of the *Gehan-Breslow test* for H_0 is

$$Z(t_0) = \int_0^{t_0} Y_1(s)Y_2(s)\left(d\hat{A}_1(s) - d\hat{A}_2(s)\right) \tag{1}$$

Here $\hat{A}_h(t)$ is the *Nelson-Aalen estimator* for $A_h(t) = \int_0^t \alpha_h(s)ds$ $(h = 1, 2)$.

(Continued on page 4.)

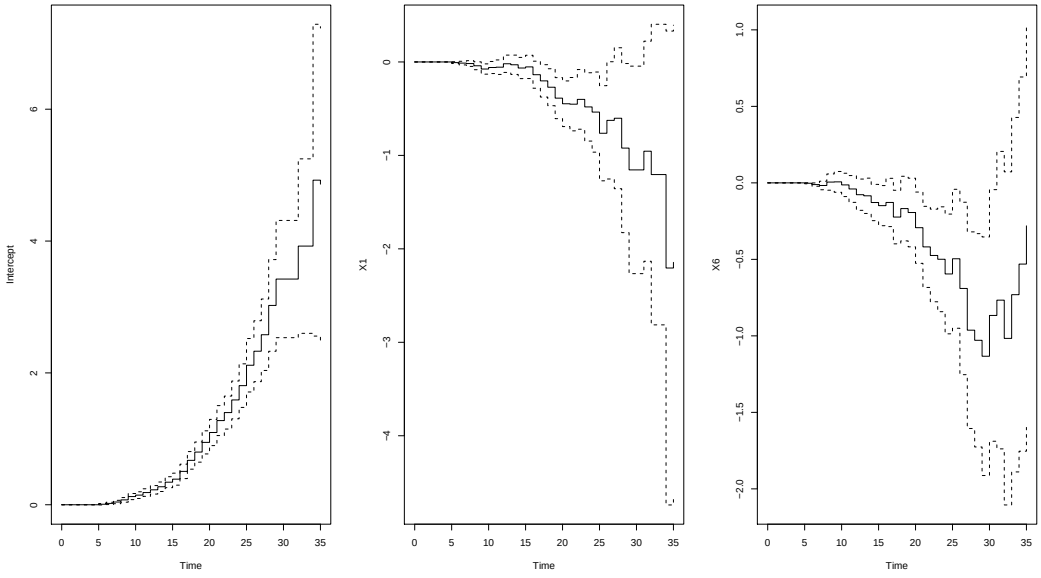

Figure 1: Plots from fitting Aalen's model to the data with covariates x_1 and x_6.

a) Looking at equation (1) and comparing to the corresponding expression for the *log-rank* test, how would you describe the practical difference between the two tests? (You should be brief here).

Show that the test statistic $Z(t_0)$ can be written

$$Z(t_0) = \int_0^{t_0} Y_2(s)\, dN_1(s) - \int_0^{t_0} Y_1(s)\, dN_2(s)$$

b) Show that under H_0 we have

$$Z(t_0) = \int_0^{t_0} Y_2(s)\, dM_1(s) - \int_0^{t_0} Y_1(s)\, dM_2(s)$$

where $M_1(t)$ and $M_2(t)$ are mean zero martingales.

How are the martingales $M_1(t)$ and $M_2(t)$ defined here?

c) Assume that H_0 holds.

Show that $Z(t)$ is a mean zero martingale (as a function of t) with predictable variation

$$\langle Z \rangle (t) = \int_0^t Y_1(s) Y_2(s) Y_\bullet(s) \alpha(s) ds \tag{2}$$

(*Hint:* You may use without proof that for counting process martingales M_1, M_2 we have $\langle M_1 + M_2 \rangle (t) = \langle M_1 \rangle (t) + \langle M_2 \rangle (t)$).

d) Use equation (2) to derive an unbiased estimator $V(t_0)$ for $\mathrm{Var}(Z(t_0))$ under H_0.

(*Hint:* You may use that $N_\bullet(t) = \int_0^t Y_\bullet(s)\alpha(s)ds + M_\bullet(t)$ under H_0.)

It can be shown (you are not asked to do this) that $Z(t_0)$ is approximately normally distributed with expected value 0 when H_0 holds and the number of individuals in the study is large.

Use this to suggest a test statistic for H_0 and determine its distribution under H_0.

(Continued on page 5.)

Problem 3 Poisson processes with unobserved heterogeneity

Suppose that n independent homogeneous Poisson processes

$$N_1(t), \ldots, N_n(t)$$

are under study, where the ith process, $N_i(t)$, has constant rate $\theta_i > 0$ and is observed in the time interval $[0, \tau_i]$ for fixed $\tau_i > 0$ $(i = 1, 2, \ldots, n)$.

a) What is in general meant by the *intensity* of a counting process $N(t)$?

Explain that the intensity of the process $N_i(t)$ is

$$\lambda_i(t) = I(t \leq \tau_i)\theta_i; \quad t \geq 0, \quad i = 1, \ldots, n$$

Then use the general expression for a parametric likelihood function (as derived in the course) to explain that the likelihood contribution for the the process $N_i(t)$ can be written

$$\theta_i^{N_i(\tau_i)} \exp\{-\theta_i \tau_i\}$$

Suppose now that each process $N_i(t)$ has a separate unobserved frailty variable Z_i such that, conditional on Z_i, the process $N_i(t)$ is a homogeneous Poisson process with rate given by

$$\theta_i = Z_i \theta$$

Here $\theta > 0$ is a common basic rate for the n processes, while $Z_1, \ldots, Z_n$ are independent and gamma-distributed with expected value 1 and variance $\delta > 0$.

The definition and some results for the gamma-distribution are given at the end of this Problem.

b) Find an expression for the *unconditional* likelihood contribution for the process $N_i(t)$.

c) Show that the log-likelihood of the data from all the n processes can be written

$$\begin{aligned} \ell(\theta, \delta) &= N \log \theta + \sum_{i=1}^{n} \log \left\{ \frac{\Gamma\left(N_i(\tau_i) + \delta^{-1}\right)}{\Gamma(\delta^{-1})} \right\} \\ &- n\delta^{-1} \log \delta - \sum_{i=1}^{n} (N_i(\tau_i) + \delta^{-1}) \log(\theta \tau_i + \delta^{-1}) \end{aligned}$$

where $N = \sum_{i=1}^{n} N_i(\tau_i)$.

Derive the score function $U_1(\theta, \delta) = \frac{\partial}{\partial \theta} \ell(\theta, \delta)$.

Show that the maximum likelihood estimator $\hat{\theta}$ for θ can be found from this score function alone if the τ_i are all equal, i.e., $\tau_i = \tau$ for $i = 1, \ldots, n$.

Find an expression for $\hat{\theta}$ in this case, and give it a suitable interpretation.

How would you next find the maximum likelihood estimator for δ? (You need not do any calculations here).

(Continued on page 6.)

d) For the *shared frailty model*, studied in the course, one considers how to recover (i.e., "estimate") the value of the frailty Z_i for each cluster i.

How would you, in the present Problem, compute recovered values for Z_i for $i = 1, 2, \ldots, n$ based on the data from all the n processes? Find explicit expressions for the recovered Z_i.

The gamma distribution

If $Z \sim \text{gamma}(k, \gamma)$, for $k, \gamma > 0$, then

$$\begin{aligned} f_Z(t) &= \frac{\gamma^k}{\Gamma(k)} t^{k-1} e^{-\gamma t} \text{ for } t > 0 \\ E(Z) &= \frac{k}{\gamma} \\ Var(Z) &= \frac{k}{\gamma^2} \\ E\left[Z^a e^{-bZ}\right] &= \frac{\Gamma(a+k)}{\Gamma(k)} \frac{\gamma^k}{(b+\gamma)^{a+k}} \text{ for } a, b \geq 0 \end{aligned}$$

END