UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK4080 — Survival and event history analysis
Day of examination:	Friday 29 November 2019
Examination hours:	09.00 – 13.00
This problem set consists of 8 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 Treatments for prostatic cancer

A randomized controlled clinical trial to compare treatments for prostatic cancer was begun in 1967 by the Veteran's Administration Cooperative Urological Research Group. Two of the treatments used in the study were a placebo and 1.0 mg of diethylstilbestrol (DES), both administered daily by mouth. The time origin of the study is the date on which a patient was randomized to a treatment, and the end-point is the death of the patient from prostatic cancer. The survival times of patients who died from other causes, or who were lost during the follow-up process, were regarded as censored.

The data used in this exercise are from a certain subset of the patients, with cancer having reached a certain stage, see the last page of the exam for these data.

The following variables are recorded for each patient:

Treat: *Treatment group.* 0 = placebo; 1 = DES.

Time: *Survival time*, possibly censored (months).

Status: *Censoring status.* 0 = censored; 1 = dead.

Age: *Age* (years).

Shb: *Serum haemoglobin level* (gm/100 ml).

Size: *Size of tumour* (cm^2).

Index: *Gleason index* (a combined index of tumour stage and grade; the more advanced the tumour, the greater the value of the index).

(Continued on page 2.)

a) Use the data (given on the last page of the exam) to compute *by hand* the *Kaplan-Meier estimator* separately for the placebo patients (`Treat=0`) and the patients that received DES (`Treat=1`), without taking the other covariates into account. (Note that the DES group has only one observed death).

Draw the two curves in the same figure (a rough plot is sufficient).

Which preliminary conclusion can be drawn from this figure?

What is the estimated median survival time for the placebo patients?

Why can't the median survival time be estimated for the DES patients?

What is meant by the *restricted mean survival time*? Show how the Kaplan-Meier plots can be used to compute the restricted means at 5 years (= 60 months) for the two treatment groups. You need not do the full computation.

b) Write down a *Cox regression model* for the given data including only `Treat` as a covariate.

The following is an edited output from R from this model.

```
Call:
coxph(formula = Surv(Time, Status == 1) ~ Treat, data = prosdata)

  n= 38, number of events= 6

         coef exp(coef) se(coef)      z Pr(>|z|)
Treat -1.9780    0.1384   1.0982 -1.801   0.0717 .

Concordance= 0.71  (se = 0.079 )
Likelihood ratio test= 4.55  on 1 df,   p=0.03
Wald test            = 3.24  on 1 df,   p=0.07
Score (log-rank) test = 4.42  on 1 df,   p=0.04
```

To what extent does the output show a significant effect of using DES?

What is the estimated hazard ratio for the DES treatment? How do you interpret this value in words?

Calculate an approximate 95% confidence interval for this hazard ratio.

c) Since the four prognostic variables `Age`, `Shb`, `Size` and `Index` may have an effect on the survival, one also performed a Cox-regression using these covariates in addition to the `Treat` variable. The resulting (edited) output from R is as follows:

```
Call:
coxph(formula = Surv(Time, Status == 1) ~ Treat + Age + Shb +
    Size + Index, data = prosdata)

  n= 38, number of events= 6
```

(Continued on page 3.)

```
          coef exp(coef) se(coef)      z Pr(>|z|)
Treat -1.18206   0.30665  1.21030 -0.977   0.3287
Age    0.04397   1.04495  0.07201  0.611   0.5414
Shb   -0.02213   0.97811  0.45273 -0.049   0.9610
Size   0.09397   1.09852  0.05209  1.804   0.0712
Index  0.72343   2.06149  0.34996  2.067   0.0387
```

Write down the model behind this output. Use the covariate names given in the output.

Give an interpretation of the estimated regression coefficients with respect to the influence of the corresponding covariate on the survival time of a patient.

Which of the explanatory variables have significant effect? (Use significance level 5% when investigating significance).

Does this output show a significant effect of using DES? Compare the present R-output with the one obtained in subproblem **b)** and comment.

Problem 2 Counting processes

Let $\{N(t); t \geq 0\}$ be a *counting process* adapted to a history $\mathcal{F}_t$, with intensity function $\lambda(t)$ defined by

$$\lambda(t)dt = P(dN(t) = 1|\mathcal{F}_{t-}).$$

The Doob-Meyer decomposition of the counting process $N(t)$ can be written

$$N(t) = \Lambda(t) + M(t), \tag{1}$$

where $\Lambda(t) = \int_0^t \lambda(s)ds$ is the *compensator* of $N(t)$ and $M(t)$ is a *mean zero martingale*.

a) What are the characteristic properties of a counting process $N(t)$?

What are the main properties of the compensator $\Lambda(t)$? How can it be expressed in terms of the intensity $\lambda(t)$?

What is the defining property of the martingale $M(t)$?

b) Define *the predictable variation process* $\langle M \rangle (t)$ of the martingale $M(t)$.

The following two properties hold for processes $\langle M \rangle (t)$ (you shall not prove these):

(i) $(M(t))^2 - \langle M \rangle (t)$ is a martingale (general property),

(ii) $\langle M \rangle (t) = \Lambda(t)$ when $M(t)$ is a counting process martingale as defined in (1).

Use the two properties (i) and (ii) to show that for the counting process in (1) we have

$$\mathrm{Var}(M(t)) = \mathrm{E}(\Lambda(t)).$$

(Continued on page 4.)

c) Assume that $N(t)$ is a nonhomogeneous Poisson process (NHPP) with rate function $\alpha(t)$. This means that, when defining $A(t) = \int_0^t \alpha(s)ds$,

(I) $N(t) - N(s) \sim$ Poisson$(A(t) - A(s))$ when $s < t$,

(II) $N(t) - N(s)$ is independent of $\mathcal{F}_s$ when $s < t$.

Use the properties (I) and (II) above to show that

$$N(t) - A(t)$$

is a mean zero martingale.

Use this to identify the compensator $\Lambda(t)$ and the martingale $M(t)$ in the Doob-Meyer decomposition (1) for a nonhomogeneous Poisson process $N(t)$.

(*Hint:* You need to calculate $\mathrm{E}(N(t) - A(t)|\mathcal{F}_s)$ for $s < t$.)

Problem 3 Parametric Poisson processes

Consider counting processes $N_i(t)$ for $i = 1, 2, \ldots, n$ that count occurrences of an event of interest for n individuals in the time interval $[0, \tau]$. Assume that the corresponding intensity processes are given on parametric form,

$$\lambda_i(t; \boldsymbol{\theta}); \quad i = 1, 2, \ldots, n,$$

where $\boldsymbol{\theta}$ is a vector of parameters.

The following expression for the likelihood of the parameter $\boldsymbol{\theta}$ is derived in the course book (you shall not do this):

$$L(\boldsymbol{\theta}) = \left\{ \prod_{i=1}^{n} \prod_{0<t\leq\tau} \lambda_i(t; \boldsymbol{\theta})^{\Delta N_i(t)} \right\} \cdot \exp\left\{ -\int_0^\tau \lambda_\bullet(t; \boldsymbol{\theta}) dt \right\}. \qquad (2)$$

Here $\Delta N_i(t)$ is the jump of $N_i(t)$ at time t, while $\lambda_\bullet(t; \boldsymbol{\theta}) = \sum_{i=1}^n \lambda_i(t; \boldsymbol{\theta})$ is the intensity of the cumulative process $N_\bullet(t) = \sum_{i=1}^n N_i(t)$.

a) Suppose that n independent nonhomogeneous Poisson processes with the same rate function $\alpha(t; \boldsymbol{\theta})$ are under study, where the ith process, $N_i(t)$, is observed in the time interval $[0, \tau_i]$ for fixed times $\tau_i > 0$, with events recorded at times $T_{i1}, \ldots, T_{iN_i}$.

Verify that the intensity of the ith process can be written

$$\lambda_i(t; \boldsymbol{\theta}) = I(t \leq \tau_i)\alpha(t; \boldsymbol{\theta}).$$

Then show that the likelihood (2) can be written

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \left(\prod_{k=1}^{N_i(\tau_i)} \alpha(T_{ik}; \boldsymbol{\theta}) \right) \exp\left\{ -\int_0^{\tau_i} \alpha(t; \boldsymbol{\theta}) dt \right\} \right\} \qquad (3)$$

(letting τ in (2) equal $\max_i \tau_i$).

(Continued on page 5.)

b) Assume now that the rate functions $\alpha(t;\boldsymbol{\theta})$ are given on so-called *power law* form,

$$\alpha(t;a,b) = abt^{b-1},$$

where $a > 0, b > 0$ are unknown parameters.

Show that the log-likelihood for the parameters a, b can be written

$$\ell(a,b) = N\log a + N\log b + (b-1)S - a\sum_{i=1}^{n}\tau_i^b,$$

where $N = \sum_{i=1}^{n} N_i(\tau_i)$, $S = \sum_{i=1}^{n}\sum_{k=1}^{N_i(\tau_i)} \log T_{ik}$.

Derive the score functions $U_1(a,b) = \frac{\partial}{\partial a}\ell(a,b)$ and $U_2(a,b) = \frac{\partial}{\partial b}\ell(a,b)$.

Show that explicit expressions for the maximum likelihood estimators $\hat{a}$ and $\hat{b}$ can be obtained if the τ_i have the common value τ (this is assumed only in this subproblem). Write down the expressions for these estimators.

c) Derive the *observed* information matrix for the model considered in subproblem **b)**.

Also, derive the *expected* information matrix.

Explain briefly how these matrices can be used to estimate the standard errors of the maximum likelihood estimates $\hat{a}$ and $\hat{b}$.

d) Go back to subproblem **a)**. Suppose that the events of the processes $N_i(t)$ are failures of certain repairable equipment of the same type, but that the equipment are used under possibly different conditions.

Explain briefly why a *shared frailty model* may be appropriate. What are the clusters in this case?

Consider the model where, for the ith process, the conditional rate function given the frailty Z_i is

$$Z_i\,\alpha(t;\boldsymbol{\theta}),$$

where the Z_i are independent and identically distributed with expected value 1 and variance $\delta \geq 0$.

Write down an expression for the full unconditional likelihood function for this shared frailty model.

(*Hint:* Modify the likelihood (3) by introducing the Z_i, and find expressions for the unconditional contributions from each process. Recall the Laplace transform where

$$\begin{aligned}\mathcal{L}(c) &= \mathrm{E}\{\exp(-cZ)\},\\ \mathcal{L}^{(r)}(c) &= (-1)^r\mathrm{E}\{Z^r\exp(-cZ)\}.\quad)\end{aligned}$$

(Continued on page 6.)

Problem 4 The log-rank test

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

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

We want to test the null hypothesis

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

where t_0 is the upper time limit of the study. The common (but unknown) version of the $\alpha_h(t)$ under H_0 will below be called $\alpha(t)$. We shall also let a subscript $\bullet$ mean the sum over 1 and 2.

The test statistic of the *log-rank test* is based on (you shall not derive this)

$$Z_1(t_0) = N_1(t_0) - \int_0^{t_0} \frac{Y_1(s)}{Y_\bullet(s)} dN_\bullet(s) \ \equiv \ O_1 - E_1, \tag{4}$$

(where there is a corresponding expression involving $N_2(t)$).

a) Show that $Z_1(t)$ is a mean zero martingale (as a function of t) when the null hypothesis holds.

(*Hint:* Under the null hypothesis we have

$$\begin{aligned} dN_h(s) &= Y_h(s)\alpha(s)ds + dM_h(s) \text{ for } h = 1, 2, \text{ and hence} \\ dN_\bullet(s) &= Y_\bullet(s)\alpha(s)ds + dM_\bullet(s). \ ) \end{aligned}$$

b) It can be shown (you shall not do this) that the predictable variation of $Z_1(t)$ is given by

$$\langle Z_1 \rangle (t) = \int_0^t \frac{Y_1(s)Y_2(s)}{Y_\bullet(s)} \alpha(s) ds.$$

Verify that an unbiased estimator of $\mathrm{Var}(Z_1(t_0))$ under H_0 is given by

$$V_{11}(t_0) = \int_0^{t_0} \frac{Y_1(s)Y_2(s)}{(Y_\bullet(s))^2} dN_\bullet(s). \tag{5}$$

Explain why the following is a natural test statistic for testing the null hypothesis H_0,

$$X^2(t_0) = \frac{(Z_1(t_0))^2}{V_{11}(t_0)}.$$

What is its distribution under the null hypothesis? (You need not prove this).

c) Consider the prostatic cancer data of **Problem 1**. You shall use the log-rank test in order to compare the survival distributions of the placebo group and the DES group. Let $N_1(t)$ and $N_2(t)$ be the processes counting events in, respectively, the group with `Treat`=0 and `Treat`=1.

Show how $Z_1(t_0)$ and $V_{11}(t_0)$ can be calculated from the data using (4) and (5) with $t_0 = 70$. You need not do the full calculation, but

(Continued on page 7.)

you should indicate clearly how this can be done by calculating the contributions for at least the first two event times of $N_\bullet(t)$.

You may then use that $O_1 - E_1 = 5 - 2.475 = 2.525$ and $V_{11}(70) = 1.442$ in order to calculate the test statistic $X^2(70)$.

Compare the calculated value to the 0.95 fractile of the χ^2-distribution with one degree of freedom, which is 3.84. Give a brief discussion of what you can conclude.

A slightly conservative version of the log-rank test is based on the test statistic

$$\frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}. \tag{6}$$

Calculate also this statistic and compare to $X^2(70)$. (Recall that $O_2 - E_2 = -(O_1 - E_1)$).

Compare the above results to the R-output of **Problem 1 b)**. What is the connection between the log-rank test and the Cox-model used in **Problem 1 b)**? (No derivations are asked for here).

(Continued on page 8.)

The prostatic cancer data

Row	Treat	Time	Status	Age	Shb	Size	Index
1	0	2	0	76	10,7	8	9
2	0	14	1	73	12,4	18	11
3	0	23	0	68	12,5	2	8
4	0	24	0	71	13,7	10	9
5	0	26	1	72	15,3	37	11
6	0	36	1	72	16,4	4	9
7	0	42	1	57	13,9	24	12
8	0	43	0	60	13,6	7	9
9	0	51	0	61	13,5	8	8
10	0	52	0	73	11,7	5	9
11	0	58	0	64	16,2	6	9
12	0	59	0	77	12,0	7	10
13	0	61	0	75	13,7	10	12
14	0	62	0	63	13,2	3	8
15	0	65	0	67	13,4	34	8
16	0	67	0	70	14,7	7	9
17	0	67	0	71	15,6	8	8
18	0	69	1	60	16,1	26	9
19	1	5	0	74	15,1	3	9
20	1	16	0	73	13,8	8	9
21	1	28	0	75	13,7	19	10
22	1	45	0	72	11,0	4	8
23	1	50	1	68	12,0	20	11
24	1	51	0	65	14,1	21	9
25	1	51	0	65	14,4	10	9
26	1	54	0	51	15,8	7	8
27	1	55	0	74	14,3	7	10
28	1	57	0	72	14,6	8	10
29	1	60	0	77	15,6	3	8
30	1	61	0	60	14,6	4	10
31	1	64	0	74	14,2	4	6
32	1	65	0	51	11,8	2	6
33	1	66	0	70	16,0	8	9
34	1	66	0	70	14,5	15	11
35	1	67	0	73	13,8	7	8
36	1	68	0	71	14,5	19	9
37	1	70	0	72	13,8	3	9
38	1	70	0	71	13,6	2	10

SLUTT