UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK4080 — Survival and event history analysis
Day of examination:	Monday 20 January 2020
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 The Nelson-Aalen estimator

Let $N(t)$ be a counting process with intensity process of the form

$$\lambda(t) = \alpha(t)Y(t) \tag{1}$$

a) Under what assumptions on the functions $\alpha(t)$ and $Y(t)$ does (1) define a *multiplicative intensity model*?

Assume in the rest of this Problem that these assumptions are satisfied.

Write down the Doob-Meyer representation for $N(t)$. What is the compensator of $N(t)$ in this case?

Use the Doob-Meyer representation to motivate the Nelson-Aalen estimator

$$\hat{A}(t) = \int_0^t \frac{J(u)}{Y(u)} dN(u)$$

as an estimator of the cumulative function $A(t) = \int_0^t \alpha(u)du$. Here $J(u) = I\{Y(u) > 0\}$.

b) Let $A^*(t) = \int_0^t J(u)\alpha(u)du$.

Verify that $\hat{A}(t) - A^*(t)$ is a mean zero martingale.

In what sense is the Nelson-Aalen estimator an approximately unbiased estimator of $A(t)$?

Find expressions for both the *predictable variation* process and the *optional variation* process of the martingale $\hat{A}(t) - A^*(t)$.

Derive an estimator of the variance of the Nelson-Aalen estimator at a fixed time t by using the optional variation process. Is the estimator unbiased?

Explain also how the predictable variation can be used to derive this estimator.

(*Hint:* You may need the formulas $\langle HdM \rangle (t) = \int_0^t H^2(s)\lambda(s)ds$ and $[HdM](t) = \int_0^t H^2(s)dN(s)$.)

(Continued on page 2.)

Problem 2 Breast-cancer study

A medical study was designed to determine if female breast-cancer patients, originally classified as lymph node negative by standard light microscopy, could be more accurately classified by immunohistochemical examination of their lymph nodes.

The data for 45 female breast-cancer patients with axillary negative lymph nodes and a minimum 10-year follow-up were selected. Of the 45 patients, 9 were immunoperoxidase positive, and the remaining 36 still remained negative.

Survival times (in months) for the two groups of patients are given below (+ denotes a right censored observation):

Immunoperoxidase Negative. 19, 25, 30, 34, 37, 46, 47, 51, 56, 57, 61, 66, 67, 74, 78, 86, 122+, 123+, 130+, 130+, 133+, 134+, 136+, 141+, 143+, 148+, 151+, 152+, 153+, 154+, 156+, 162+, 164+, 165+, 182+, 189+

Immunoperoxidase Positive: 22, 23, 38, 42, 73, 77, 89, 115, 144+

We shall denote the immunoperoxidase negative group as Sample 1 and the immunoperoxidase positive group as Sample 2. Let T_h for $h = 1, 2$ be the survival time for a patient in Sample h, with hazard and cumulative hazard functions given by $\alpha_h(t)$ and $A_h(t)$, respectively.

a) Calculate *(by hand)* the Nelson-Aalen estimates $\hat{A}_1(60)$ and $\hat{A}_2(60)$ based on the given data. (It is sufficient to write down the appropriate expressions. You may use in the following that the estimates are, respectively, 0.3201 and 0.5456.)

Show similarly how to calcuate estimates of the standard deviations of the two Nelson-Aalen estimates. (You may use that these standard deviation estimates are, respectively, 0.1017 and 0.2760.)

b) One wants estimates and 95% confidence intervals for the probability of 5 years survival for the patients in each group, i.e., $S_h(60)$, for $h = 1, 2$, where $S_h(t)$ is the survival function of T_h. Use results from subproblem **a)** to find both the estimates and the confidence intervals.

Problem 3 Aalen's additive model

Assume that we have counting processes $N_1(t), N_2(t), \ldots, N_n(t)$ with no simultaneous jumps, that register the occurrences of an event of interest for n individuals. For individual i ($i = 1, 2, \ldots, n$), we have a single covariate, x_i, and we assume that the intensity process of $N_i(t)$ takes the form

$$\lambda_i(t) = Y_i(t)\{\beta_0(t) + \beta_1(t)x_i\}$$

Here $Y_i(t) = 1$ if individual i is at risk "just before" time t and $Y_i(t) = 0$ otherwise. We introduce the vector

$$\mathbf{N}(t) = (N_1(t), N_2(t), \ldots, N_n(t))^T$$

(Continued on page 3.)

where ^{T} denotes transpose, and the matrix

$$\mathbf{X}(t) = \begin{pmatrix} Y_1(t) & Y_1(t)x_1 \\ Y_2(t) & Y_2(t)x_2 \\ \vdots & \vdots \\ Y_n(t) & Y_n(t)x_n \end{pmatrix}$$

We further introduce $\mathbf{B}(t) = (B_0(t), B_1(t))^T$, where $B_j(t) = \int_0^t \beta_j(u)du$ for $j = 0, 1$. From the course we know that $\mathbf{B}(t)$ may be estimated by

$$\hat{\mathbf{B}}(t) = \int_0^t J(u) \left\{\mathbf{X}(u)^T\mathbf{X}(u)\right\}^{-1} \mathbf{X}(u)^T d\mathbf{N}(u) \tag{2}$$

where $J(u) = I\{\mathbf{X}(u) \text{ has full rank}\}$.

We also introduce $\mathbf{B}^*(t) = \int_0^t J(u)d\mathbf{B}(u)$.

a) Show that $\hat{\mathbf{B}}(t) - \mathbf{B}^*(t)$ equals a vector-valued stochastic integral, and explain why this implies that $\hat{\mathbf{B}}(t)$ is almost an unbiased estimator of $\mathbf{B}(t)$.
(*Hint:* Verify first the relation $d\mathbf{N}(t) = \mathbf{X}(t)d\mathbf{B}(t) + d\mathbf{M}(t)$.)

In the following we assume that the x_i take the values 0 and 1. The individuals i with $x_i = 0$ then constitute what we shall call Sample 1, while the individuals with $x_i = 1$ constitute Sample 2.

b) Show that

$$\mathbf{X}(t)^T\mathbf{X}(t) = \begin{pmatrix} R_1(t) + R_2(t) & R_2(t) \\ R_2(t) & R_2(t) \end{pmatrix}$$

where $R_h(t)$ is the number at risk at time t in Sample h, for $h = 1, 2$.
How can the event $\{J(t) = 1\}$ be expressed by the $R_h(t)$?

In the following you may use without proof that if $J(t) = 1$, we have

$$\left(\mathbf{X}(t)^T\mathbf{X}(t)\right)^{-1} = \begin{pmatrix} \frac{1}{R_1(t)} & -\frac{1}{R_1(t)} \\ -\frac{1}{R_1(t)} & \frac{1}{R_1(t)} + \frac{1}{R_2(t)} \end{pmatrix}$$

Let $t^* = \sup\{t : J(t) = 1\}$ and let $\{T_j\}$ be the observed event times for the n processes

c) Show that for $t \le t^*$ we have

$$\begin{aligned} \hat{B}_0(t) &= \sum_{T_j \le t,\, j \in \text{Sample 1}} \frac{1}{R_1(T_j)} \\ \hat{B}_1(t) &= \sum_{T_j \le t,\, j \in \text{Sample 2}} \frac{1}{R_2(T_j)} - \sum_{T_j \le t,\, j \in \text{Sample 1}} \frac{1}{R_1(T_j)} \end{aligned}$$

(*Hint:* Considering (2), it may be useful to calculate $\mathbf{X}(u)^T d\mathbf{N}(u)$ separately.)

(Continued on page 4.)

d) Verify that the estimator $\hat{B}_0(t)$ equals the Nelson-Aalen estimator obtained using data from Sample 1 only. How can you similarly characterize $\hat{B}_1(t)$?

Why are these properties of the estimators $\hat{B}_0(t)$ and $\hat{B}_1(t)$ reasonable from the model description in the beginning of the Problem?

Use the connections to the Nelson-Aalen estimator to suggest estimators for $\mathrm{Var}(\hat{B}_0(t))$, $\mathrm{Var}(\hat{B}_1(t))$ and $\mathrm{Cov}(\hat{B}_0(t), \hat{B}_1(t))$.

Consider again the breast-cancer study from **Problem 2**. Consider the total sample of $36 + 9 = 45$ patients as a sample with a single covariate x_i, where the individuals in the immunoperoxidase negative group (Sample 1) have $x_i = 0$, while the patients in the immunoperoxidase positive group (Sample 2) have $x_i = 1$.

The data were analysed by the function `aareg` in R, using the following commands:

```
> fit.aalen=aareg(Surv(T,cens==1)~x)
> print(fit.aalen)
> par(mfrow=c(1,2))
> plot(fit.aalen)
```

The output of the `print()` command is as follows:

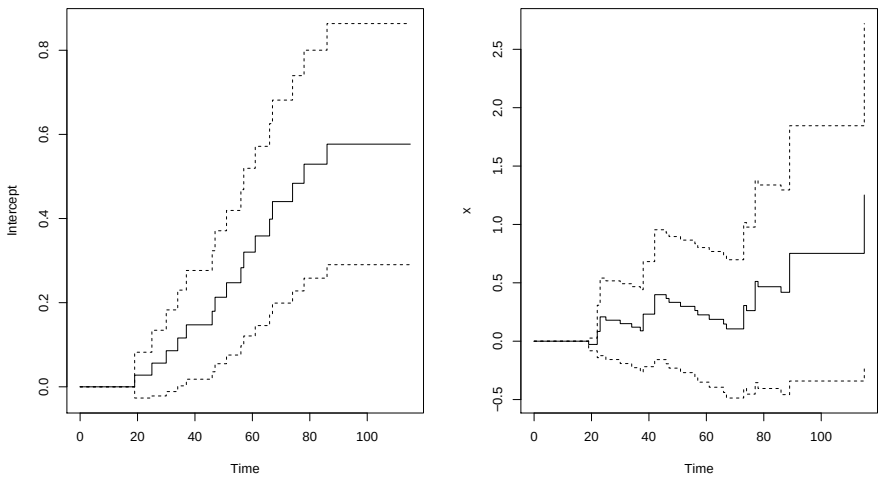
```
Call:
aareg(formula = Surv(T, cens == 1) ~ x)

  n= 45
    24 out of 24 unique event times used

            slope   coef se(coef)   z        p
Intercept 0.00687 0.0238  0.00594 4.0 6.33e-05
x         0.00909 0.0384  0.02250 1.7 8.84e-02

Chisq=2.9 on 1 df, p=0.0884; test weights=aalen
```

while the plot from the `plot()` command is given below:



(Continued on page 5.)

e) Explain briefly what you can conclude from these results. What is t^* for these data? What can you say about the difference in hazard between the Immunoperoxidase Negative and the Immunoperoxidase Positive group?

Problem 4 The proportional frailty model

In this exercise you may use the following properties of the gamma-distribution:

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

$$\begin{aligned} f_Z(t) &= \frac{\gamma^k}{\Gamma(k)} t^{k-1} e^{-\gamma t} \text{ for } t > 0 \\ \mathcal{L}(c) &\equiv E(e^{-cZ}) = \left(1 + \frac{c}{\gamma}\right)^{-k} \\ E(Z) &= \frac{k}{\gamma} \\ Var(Z) &= \frac{k}{\gamma^2} \\ E(Z^{-1}) &= \frac{\gamma}{k-1} \quad \text{for } k > 1 \\ E(Z^{-2}) &= \frac{\gamma^2}{(k-1)(k-2)} \quad \text{for } k > 2 \end{aligned}$$

Assume that the hazard rate of an individual's lifetime, T, conditional on a frailty, Z, is expressed as the product

$$Z\, \alpha(t)$$

Here $\alpha(t)$ is a basic hazard rate, while the frailty Z is assumed to be gamma-distributed with expected value 1 and variance $\delta > 0$, i.e., $Z \sim \text{gamma}(1/\delta, 1/\delta)$.

a) Explain briefly how the proportional frailty model is used to model unobserved heterogeneity within a population.

In the setting defined above, what is meant by the population distribution for T? Show that the population survival function in the given situation can be expressed as

$$S(t) = (1 + \delta A(t))^{-\frac{1}{\delta}}$$

b) Assume *in this subproblem* that $\alpha(t) = \theta > 0$ for all $t \geq 0$.

Find expressions for the population survival function, the population hazard, the population mean and the population variance as functions of δ and θ.

Discuss the form of these expressions in the given situation.

(Continued on page 6.)

c) Show that the conditional distribution of Z given $T = t$ is

$$\text{gamma}\left(1 + \frac{1}{\delta},\ A(t) + \frac{1}{\delta}\right)$$

(Hint: Use Bayes' formula: "$f(z|t) \propto f(t|z)f(z)$" .)

Use this to find expressions for $E(Z|T = t)$ and $Var(Z|T = t)$ in terms of δ and $A(t)$.

How do you interpret these results?