UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK4080/STK9080 — Survival and event history analysis
Day of examination:	Friday December 9th, 2016
Examination hours:	09.00 – 13.00
This problem set consists of 4 pages.	
Appendices:	None
Permitted aids:	Appproved calculator.

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

Problem 1

Let $T_i, i = 1, \ldots, n$ be iid with survival function $S(t) = \mathrm{P}(T_i > t)$. Assume that we observe right censored data $\tilde{T}_i = \min(c_i, T_i)$ and $D_i = I(\tilde{T}_i = T_i)$ where the censoring times $c_1, \ldots, c_n$ are given numbers.

a

Write down an expression for the Kaplan–Meier estimator $\hat{S}(t)$ of $S(t)$ and give an intuitive explanation for this estimator.

b

Make a sketch of a Kaplan–Meier estimator with pointwise confidence intervals. (You are not asked to explain how the confidence intervals are calculated).

Demonstrate how you can estimate percentiles with confidence intervals from such plots.

c

For survival times with $\mathrm{P}(T_i \geq 0) = 1$ we have the general representation $\mu = \mathrm{E}[T_i] = \int_0^\infty S(t)dt$ (You are not asked to show this relation).

A natural estimator for μ from right censored survival data is given by $\hat{\mu} = \int_0^\infty \hat{S}(t)dt$. Discuss problems with this estimator.

Show that for uncensored survival data, i.e. all $\tilde{T}_i = T_i$, the estimator becomes $\hat{\mu} = \overline{T} = \frac{1}{n}\sum_{i=1}^n T_i$.

(Continued on page 2.)

Problem 2

Assume iid T_i with hazard $\alpha(t;\theta) = \exp(\theta)$, so the T_i are exponentially distributed. We observe, as in Problem 1, right censored data $(\tilde{T}_i, D_i)$ where $\tilde{T}_i = \min(T_i, c_i)$, $D_i = I(\tilde{T}_i = T_i)$ and the c_i are given numbers.

a

Argue that the log-likelihood of the $(\tilde{T}_i, D_i), i = 1, \ldots, n$ can be written as $l(\theta) = \sum_{i=1}^n l_i(\theta)$ where

$$l_i(\theta) = \theta D_i - \exp(\theta)\tilde{T}_i.$$

Show that the maximum likelihood estimator of θ is given by $\hat{\theta} = \log(D_\bullet/R_\bullet)$ where $D_\bullet = \sum_{i=1}^n D_i$ and $R_\bullet = \sum_{i=1}^n \tilde{T}_i$.

b

Let $\tau = \max(c_1, \ldots, c_n)$ be the maximum potential follow-up time. Show that the score contributions, i.e. derivatives of $l_i(\theta)$ with respect to θ, can be written as

$$u_i(\theta) = N_i(\tau) - \int_0^\tau Y_i(s)\exp(\theta)ds$$

where $N_i(t) = I(\tilde{T}_i \leq t, D_i = 1)$ and $Y_i(t) = I(\tilde{T}_i \geq t)$.

Use martingale theory to show that the score contributions have $\mathrm{E}[u_i(\theta)] = 0$ and derive an expression for their variances.

Furthermore, state the expectation of the full score function $u(\theta) = \sum_{i=1}^n u_i(\theta)$ and give an expression for the variance of $u(\theta)$.

c

Find an expression for the observed information $I(\theta) = -\frac{\partial u(\theta)}{\partial \theta}$ and demonstrate that $\mathrm{E}[I(\theta)] = \mathrm{Var}[u(\theta)]$.

Suggest an estimator for the variance of $\hat{\theta}$ and give an approximate 95% confidence interval for the hazard rate $\alpha(t;\theta) = \exp(\theta)$.

Problem 3

The Bergen Clinical Blood Pressure Survey examined a sample men and women in the city of Bergen, Norway, in 1965–71. This sample was followed until 2006 with respect to emigration and mortality. Here we investigate the effect on total mortality of covariates sex ($x_{i1} = 0$ for men and $x_{i1} = 1$ for women), physical activity ($x_{i2} = 0$ for no/low activity and $x_{i2} = 1$ for any activity over "low") and cholesterol ($x_{i3} = 0$ for below median and $x_{i3} = 1$ for above median). The analyses are restricted to the $n = 2382$ participant aged less than 45 years at initial examination and use attained age as time variable.

(Continued on page 3.)

a

State a Cox-model for these data. Explain how the model can be fitted without having to specify the baseline hazard.

Interpret the results from the following output with a Cox-regression model. In particular calculate approximate 95% confidence intervals for the hazards ratios.

```
                coef exp(coef) se(coef)     z       p
sex          -0.6543    0.5198   0.0751 -8.72 < 2e-16
activity     -0.3391    0.7124   0.0879 -3.86 0.00011
cholesterol   0.2999    1.3497   0.0747  4.02 5.9e-05

Likelihood ratio test=105  on 3 df, p=0
n= 2382, number of events= 743
```

b

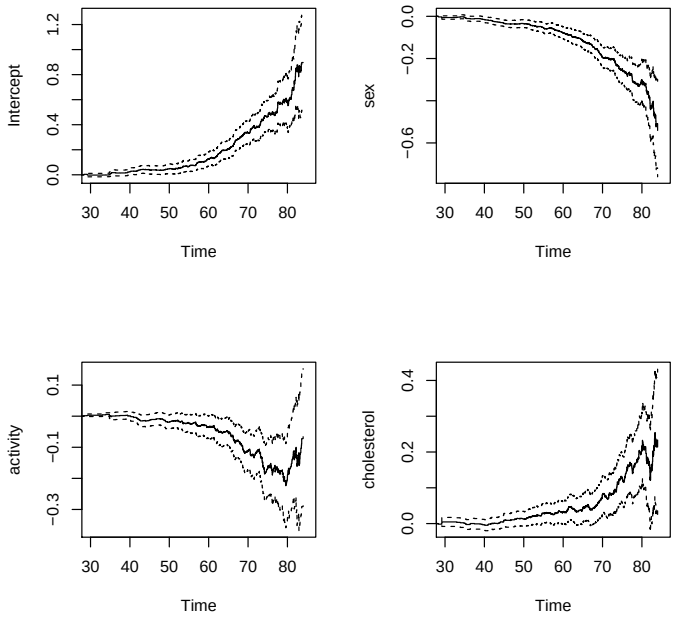
Assume that the proportional hazards assumption of the Cox-regression holds for the two covariates x_{i2} physical activity and x_{i3} cholesterol, but fails for the covariate x_{i1} sex. Present the concept of stratified Cox-regression and explain why it can be used as a remedy in this case.

How would you present (display) differences in mortality between men and women in such a situation? Explain how an appropriate estimator for showing such differences is calculated.

c

Another options when the proportional hazards assumption fail is to use the Aalen additive hazards model. State this model.

Below you see a plot from fitting the Aalen additive hazards model with the same covariates as in question a). The plots give cumulative regression functions. Explain how the plots should be interpreted.



(Continued on page 4.)

d

Explain how the cumulative regression functions like those in the previous question are calculated, in particular note which estimating method is used and specify how the response variables and covariates are set up.

Give a brief explanation for why the estimator of the cumulative regression functions is close to unbiased.

END