Exam: Survival analysis

1 Theoretical Part

1.1 Part A

Let

$$N^g(t) = (N_i^g(t); i = 1, \ldots, n_g)$$

be a multivariate counting process with filtration F_t^g and F_t^g intensity given by

$$\lambda_i^g(t) = \alpha_i^g(t, X_i^g(t), Z_i^g(t))Y_i^g(t),$$

where $Y_i^g(t)$ is a non-negative predictable process (think of it as the at risk indicator), $Z_i^g(t)$ is a p-vector and $X_i^g(t)$ a q-vector of predictable and bounded covariates, for $g = f, m$. Females (f) and males (m). All subjects are independent within the groups, and in addition the groups of males and females are independent.
The counting processes are related to survival times T_i^g such that the hazard of T_i^g is given by

$$\alpha_i^g(t, X_i^g(t), Z_i^g(t)).$$

Males and females are recruited to our study when they are 30 years old and followed in up to 60 years until death or censoring. We are interested in relating the risk of dying to $Z_i^g(t)$ and $X_i^g(t)$ out covariates. Let further $\Lambda_i^g(t) = \int_0^t \lambda_i^g(s)\, ds$ be the compensator of $N_i^g(t)$ and $M_i^g(t) = N_i^g(t) - \Lambda_i^g(t)$ the counting process martingale. Let $B(t) = \int_0^t \beta(s)ds$. You may make any additional regularity conditions as needed.
We assume that the males have intensities given by

$$\lambda_i^m(t) = Y_i^m(t)((X_i^m(t))^T\beta(t) + h(\gamma^T Z_i^m(t)))$$

and that the females have intensity

$$\lambda_i^f(t) = Y_i^f(t)((X_i^f(t))^T\beta(t-4) + h(\gamma^T Z_i^f(t)))$$

such that a female is like a 4 year younger male and h is some known smooth link function such as the identity or $\exp(\cdot)$. Think of the h-term as giving some excess risk due to the risk factors given by the covariates Z.

a) Write up the hazard for a the survival time of a female plus 4 years, that is $T_i^f + 4$.

We now wish to establish estimators of the parameters of the model, that is $(B(t), \gamma)$.

b) Write up an estimating equations for the $B(t)$ given γ.

c) Write up an estimating equation for γ and using b) to profile out $B(t)$.

d) Outline the main arguments in estimating γ and how to establish that $\sqrt{n}(\hat{\gamma} - \gamma)$ is asymptotically normal under regularity conditions. Show how to estimate the variance of this asymptotic normal distribution.

d) Outline the main arguments in estimating the cumulatives, $B(t)$, and that $\sqrt{n}(\hat{B}(t) - B(t))$ is asymptotically normal with a variance that we can estimate (for fixed t). Show how to estimate the variance of this asymptotic distribution.

d) For which t can we estimate $B(t)$ based on the data.

e) What is the predicted survival probability after 10 years for a male subject with covariates (X_0, Z_0) that enters the study at 30 years of age. Indicate how to get standard errors for this estimate.

f) It turns out that the females in the study are recruited at different ages, and are all older than 30 when being recruited for the study. We know their individual ages of recruitment. How can you modify the above analysis to still estimate $B(\cdot), \gamma)$ consistently.

1.2 Part B

Now consider the following hazard

$$\lambda(t) = \beta(t) + X\alpha(t + A)$$

for $t \in [0, \tau]$ given stochastic variables X and A. X is a binary covariate where $P(X = 1) = P(X = 0) = 1/2$, and A is a positive stochastic variable with density $f(a)$. Assume that X and A are independent. We assume that this hazard is related to a survival time T that is observed with independent right censoring.

a) What is the hazard function given only X, that is when X is observed and A is not observed.

b) What is the hazard function when both X and A are not observed.

2 Practical part

We shall consider a data-set on 423 first pregnancy planners and study risk-factors that may affect the "time to pregnancy" (TTP).
Members of four nationwide unions in the age range from 20-35 living with a partner were invited to join a TTP study. Inclusion criteria were: no prior knowledge of fertility and current use of contraception but planning to discontinue within the study period in order to conceive. 50000 couples were contacted and the first 430 included in the study. We consider 423 of these. These couples were followed for 6 months or until conception was achieved. Various information was collected about the couples at the initiation of the study. Lifestyle factors, blood samples, and a semen sample from the male partner.
The data-set can be downloaded from the course homepage as an xpt SAS file and below it is described how you can get it into R.
The following variables are defined in the data-set:

`obsnr`	observation number
`ttp`	TTP in days (continuous variable)
`k_gravid`	censoring variable (pregnancies=1, censorings=0)
`f_cyklus`	number of menstrual cycles to pregnancy
`f_xid`	identification number
`k_alk`	number of drinks for the female
`k_cof`	intake of caffeine for the female (mg per day)
`m_alk`	number of drinks for the male
`m_cof`	intake of caffeine for the male (mg per day)
`k_mryg`	smoking status of the females mother
`m_mryg`	smoking status of the males mother
`k_ryg`	smoking status of the female
`m_ryg`	smoking status of the male
`m_zkon0`	sperm concentration of male (mill/ml)
`mkryg`	0 if none smokes, 1 if `m_ryg`=1 and `k_ryg`=0
`mkryg`	3 if both smokes, 2 if `m_ryg`=0 and `k_ryg`=1

Using the continuous version of the TTP (ttp) in the data set. To load into R is easy :

```
> library(foreign)
> ttp<-read.xport("ttp.xpt")
> names(ttp)<-to.lower(names(ttp))
```

Consider the following risk factors in a model :

`k_cof`	intake of caffeine for the female (mg per day)
`k_ryg`	smoking status of the female
`m_ryg`	smoking status of the male
`m_zkon0`	sperm concentration of male (mill/ml)

Build a model that describes the effect of these risk factors on the TTP. We expect that smoking and caffeine is bad for fertility and that a high sperm concentration leads to better fertility. Consider the following questions:

1) Fit a Cox regression model and do a goodness-of-fit (GOF) analyses. Check for interactions.

2) Fit an additive hazards model and do a GOF for the model. How can you summarize the effects as simply as possible. Estimate the survival function for various subgroups (with confidence bands based on resampling) and compare with the results in 1).

3) How would you summarize the findings based on the above models. Is there a clear indication that one model is superior to others ?