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

a

With $Y(t) = \sharp\{\tilde{T}_i \geq t)$ = the number at risk at time t the Kaplan-Meier estimator is given by

$$\hat{S}(t) = \prod_{\tilde{T}_i \leq t} \left[1 - \frac{D_i}{Y(\tilde{T}_i)}\right]$$

assuming that there are no ties.

The expression in the brackets $1 - \frac{D_i}{Y(\tilde{T}_i)} = \hat{\mathrm{P}}(T > \tilde{T}_i | T > \tilde{T}_i -)$ can be interpreted as the (estimated) conditional probability of surviving at time $\tilde{T}_i$ conditional on being alive at time $t-$ and so also the conditional probability of surviving from the previous event time. Then by the general rule $\mathrm{P}(A \cap B) = \mathrm{P}(A|B)\mathrm{P}(B)$ we get the Kaplan-Meier estimator as $\hat{S}(t) =$

$$\hat{\mathrm{P}}(T > t) = \hat{\mathrm{P}}(T > t | T > t_k)\hat{\mathrm{P}}(T > t_k | T > t_{k-1})\hat{\mathrm{P}}(T > t_{k-1} | T > t_{k-2}) \cdots \hat{\mathrm{P}}(T > t_1 | T > t_0)$$

where $t_0 = 0 < t_1 < t_2 < \cdots < t_k < t$ are the observed event times.

b

See the file `sketch1b.pdf`. For percentile μ_p defined by $S(\mu_p) = 1 - p$ we obtain the estimate $\hat{\mu}_p$ solving $\hat{S}(\hat{\mu}_p) = 1 - p$. Graphically we find the estimate by drawing a horizontal line through $y = 1 - p$ and reading off where this line crosses the Kaplan-Meier. Similarly a confidence interval is determined as the values $\hat{\mu}_p^L < \hat{\mu}_p^U$ where the horizontal line crosses the lines for the confidence interval of the Kaplan-Meier estimator.

c

Very often $\hat{S}(t_{\max}) > 0$ where $t_{\max}$ is the maximum of the the right censoring times. Then the estimator is not well defined for $t > t_{\max}$ and so neither is the estimator of μ.

(Continued on page 2.)

With uncensored data we get $\hat{S}(t) = 1 - \frac{k}{n}$ where k is the number of $T_i < t$. Then with $T_{(1)} < T_{(2)} < \cdots < T_{(n)}$ the ordered T_i's we get

$$\begin{aligned}\hat{\mu} &= 1 * T_{(1)} + \tfrac{n-1}{n}(T_{(2)} - T_{(1)}) + \tfrac{n-2}{n}(T_{(3)} - T_{(2)}) + \cdots + \tfrac{1}{n}(T_{(n)} - T_{(n-1)}) \\ &= T_{(1)}(1 - \tfrac{n-1}{n}) + T_{(1)}(\tfrac{n-1}{n} - \tfrac{n-2}{n}) + \cdots + T_{(n-1)}(\tfrac{2}{n} - \tfrac{1}{n}) + \tfrac{1}{n}T_{(n)} \\ &= \tfrac{1}{n}T_{(1)} + \tfrac{1}{n}T_{(2)} + \cdots + \tfrac{1}{n}T_{(n)} = \overline{T}\end{aligned}$$

Alternatively one can realize the result graphically by inspecting the diagram in the file `sketch1c.pdf` with $n = 7$. Here each (horizontal) rectangle has area $\frac{1}{n}T_{(k))}$ and the total area under the survival function becomes $\overline{T}$.

Problem 2

a

The likelihood contributions can be written as, with $f(t;\theta)$ the density and $S(t;\theta)$ the survival function of T_i

$$\begin{aligned}L_i(\theta) &= f(\tilde{T}_i;\theta)^{D_i} S(\tilde{T}_i;\theta) = \alpha(\tilde{T}_i;\theta)^{D_i} \exp(-\int_0^{\tilde{T}_i} \alpha(t;\theta)dt) \\ &= \exp(\theta D_i)\exp(-\exp(\theta)\tilde{T}_i)\end{aligned}$$

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

The score contributions $u_i(\theta)$, i.e. the derivatives of $l_i(\theta)$ with respect to θ becomes $u_i(\theta) = D_i - \tilde{T}_i \exp(\theta)$. Hence the score equals $u(\theta) = \sum_{i=1}^n D_i - \exp(\theta)\sum_{i=1}^n \tilde{T}_i = D_\bullet - \exp(\theta)R_\bullet$. Putting this equal to zero gives $\hat{\theta} = \log(D_\bullet/R_\bullet)$.

b

We note that $N_i(\tau) = I(D_i = 1, \tilde{T}_i \le \tau) = D_i$ and

$$\tilde{T}_i \exp(\theta) = \int_0^\tau I(\tilde{T}_i > t)\exp(\theta)dt = \int_0^\tau Y_i(t)\exp(\theta)dt.$$

Furthermore the intensity process of the counting process $N_i(t)$ is given by $\lambda_i(t) = Y_i(t)\alpha(t;\theta) = Y_i(t)\exp(\theta)$. This means $M_i(t) = N_i(t) - \int_0^t Y_i(s)\exp(\theta)ds$ is a martingale with expectation zero and so also $u_i(\theta) = M_i(\tau)$ has expectation zero.

Moreover the predictable variation process of $M_i(t)$ equals $\Lambda_i(t) = \int_0^t Y_i(s)\exp(\theta)ds$ and so $\mathrm{var}(u_i(\theta)) = \mathrm{E}[\Lambda_i(\tau)] = \mathrm{E}[\int_0^\tau Y_i(s)\exp(\theta)ds]$

Finally the full score $u(\theta) = M_\bullet(\tau) = \sum_{i=1}^n M_i(\tau)$, and has expectation zero and variance given by $\mathrm{E}[\int_0^\tau Y(s)\exp(\theta)ds]$ where $Y(s) = \sum_{i=1}^n Y_i(s)$.

c

We find $I(\theta) = -\frac{d}{d\theta}u(\theta) = \Lambda(\tau) = \int_0^\tau Y(s)\exp(\theta)ds$ and so from b) $\mathrm{E}[I(\theta)] = \mathrm{E}\int_0^\tau Y(s)\exp(\theta)ds = \mathrm{var}(u(\theta))$.

Inserting $\exp(\hat{\theta}) = D_\bullet/R_\bullet$ into $I(\theta)$ gives $I(\hat{\theta}) = D_\bullet = N(\tau)$. Thus we estimate $\mathrm{var}(\hat{\theta})$ by $1/I(\hat{\theta}) = 1/D_\bullet$. Furthermore an approximate 95% confidence interval for $\alpha(t;\theta) = \exp(\theta)$ is given by $(D_\bullet/R_\bullet)\exp(\pm 1.96/\sqrt{D_\bullet})$.

(Continued on page 3.)

Problem 3

a

The Cox-model is given by individual hazards $\alpha_i(t) = \alpha_0(t)\exp(\beta' x_i)$. We may fit the model by maximizing the partial likelihood

$$L(\beta) = \prod_{i:D:i=1} \frac{\exp(\beta' x_i}{\sum_{j \in \mathcal{R}(\tilde{T}_i)} \exp(\beta' x_j)}$$

which can be given the interpretation as the product over probabilities that individual i with $D_i = 1$ died at $\tilde{T}_i$ given that one of those in the risk set $\mathcal{R}(\tilde{T}_i)$ at $\tilde{T}_i$ died.

We see that all three covariates are strongly significant. Women have a hazard which estimated as being only 0.52 the hazard of men. Those who are physically active have a hazard rate being 0.71 of the non-active and people with higher cholesterol than the median has a mortality (hazard) which 1.35 time that of an below median cholesterol individual (when other factors are the same).

Approximate 95% confidence intervals for the hazard ratios are $\exp(\hat{\beta}_j + \pm 1.96 se_j)$ where the $\hat{\beta}_j$ are the estimates of β_j and se_j their standard errors. We get intervals (0.449,0.602) for sex, (0.600,0.846) for activity and (1.166,1.563) for cholesterol.

b

The idea of stratified Cox-regression is that the hazard can be written (in this case) as $\alpha_i(t) = \alpha_{0x_{i1}}(t)\exp(\beta_2 x_{i2} + \beta_3 x_{i3})$, that is with different baselines for men and women, but where log-hazard ratios β_2 and β_3 are the same for men and women. Now there is no assumption of hazard ratios between women and men being time-constant.

One may then calculate separate partial likelihood $L_M(\beta_2, \beta_3)$ for men and $L_W \beta_2, \beta_3)$ for women and combine this into a total stratified partial likelihood $L(\beta_2, \beta_3) = L_M(\beta_2, \beta_3) L_W(\beta_2, \beta_3)$.

To display the differences in risk between men and women one may calculated seperate cumulative hazards for men and women by the Breslow-estimators, possibly translated to survival functions.

These Breslow estimators of the baseline cumulative hazards can be given as

$$\hat{A}_j(t) = \int_0^t \frac{dN_{\bullet j}(s)}{S_j^{(0)}(s; \hat{\beta})}$$

where $N_{\bullet j}(s) = \sum_{i:x_{i1}=j} N_i(s)$ and $S_j^{(0)}(s; \hat{\beta}) = \sum_{i:x_{i1}=j} Y_i(s)\exp(\hat{\beta}' x_i)$. Here $N_i(s)$ and $Y_i(s)$ are individual counting processes and indicators of being at risk.

c

The additive hazards model can be written as

$$\alpha_i(t) = \beta_0(t) + \beta_1(t) x_{i1} + \beta_3(t) x_{i3} + \beta_3(t) x_{i3}$$

(Continued on page 4.)

where $\beta_0(t)$ is the hazard if all covariates are equal to zero and $\beta_j(t)$ are regression functions being positive if there is a higher risk associated with large than small x_{ij} at time t and negative in the opposite case. The effect of the covariates is thus allowed to vary with time.

The plots shows estimates of $B_j(t) = \int_0^t \beta_j(s)ds$. Thus when $\hat{B}_0(t)$ is a close to concave function this shows that the baseline hazard increases over time. Furthermore $\hat{B}_1(t)$ is decreasing ever more steeply, thus women have a smaller hazard than men and the difference increases over time. For physical activity we see a negative cumulative hazard, only slightly below zero up to 50 years, then decreasing more markedly and perhaps stabilizing towards 80 years (likely this last observation is not significant). In correspondence with the results from the Cox-regressions the hazard is smaller for the physical active, but we estimate that the difference first increases and then possibly vanishes when the individuals get into their 70's and 80's. For cholesterol the high cholesterol group consistently has a higher hazard, but the association is weak into the 60's or 70's and then increase more steeply.

d

The estimates $\hat{B}(t) = (\hat{B}_0(t), \hat{B}_2(t), \hat{B}_2(t), \hat{B}_3(t))'$ have increments $d\hat{B}(t) = (d\hat{B}_0(t), d\hat{B}_2(t), d\hat{B}_2(t), d\hat{B}_3(t))'$ at event times $\tilde{T}_i, D_i = 1$. For these times the increments are obtained as least squares estimates with the $dN_i(t)$ as responses and covariates $x_{ij}Y_i(t)$.

We can write $dN_i(t) = Y_i(t)\alpha_i(t) + dM_i(t)$ where $dM_i(t)$ are martingale increments with expectation zero. After some algebra it is then possible to show that the estimators of the estimated cumulative regression functions can be written as the true cumulative regression functions plus a martingale term with expectation zero as long as the design matrices at different event times have full rank. Thus the estimates are approximately unbiased.

(To do most of this algebra we note that the estimates $d\hat{B}(t)$ can be written as

$$d\hat{B}(t) = (X(t)^\top X(t))^{-1} X(t)^\top dN(t)$$

where $X(t)$ is the design-matrix and $dN(t)$ vector the indicators of ecents at time t. This vector of responses can be expanded to $X(t)dB(t) + dM(t)$ where $dM(t)$ is the vector of the $dM_i(t)$'s. Inserting this into the expression for $d\hat{B}(t)$ we get, as long as the design matrix has full rank,

$$d\hat{B}(t) = (X(t)^\top X(t))^{-1} X(t)^\top (X(t)dB(t) + dM(t)) = dB(t) + (X(t)^\top X(t))^{-1} X(t)^\top dM(t)$$

where the latter term is (vector of) martingale increment. It turns out that for t such that the design matrix has full rank for $s \leq t$

$$\hat{B}(t) = B(t) + M^\star(t)$$

where the last term is a vector of martingales with expectation zero.)

END