UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in:	STK4080/STK9080 — Survival and Event History Analysis.
Day of examination:	Monsday, December 11, 2023.
Examination hours:	15.00 – 19.00.
This problem set consists of 4 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.

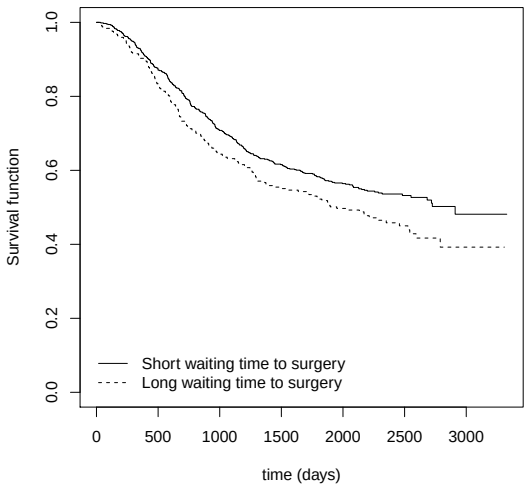
Problem 1

a) Let $\alpha(t)$ be a hazard function and $S(t)$ the corresponding survival function for a (continuously distributed) survival time T. Show the identity

$$S(t) = \exp(-\int_0^t \alpha(s)ds).$$

b) Let $Y_{\bullet j}(t)$ denote the number at risk and $N_{\bullet j}(t)$ the counting processes of events in two groups. State the formula for the Kaplan-Meier estimators of the survival functions for the groups $j = 1, 2$.

Below you see a plot of two such Kaplan-Meier estimators for time to death from a study of colon cancer where group one is those with short time and group two those with long time to surgery. Describe the difference in survival between the two groups.



(Continued on page 2.)

c) It is usually considered insufficient to just present Kaplan-Meier curves without a formal test of difference between the groups. The most commonly used test for this purpose is the log-rank-test which is based on a test-statistic, for follow-up on an interval $[0, \tau]$,

$$Z(\tau) = N_{\bullet 2}(\tau) - E_2(\tau) = N_{\bullet 2}(\tau) - \int_0^\tau \frac{Y_{\bullet 2}(t)}{Y_\bullet(t)} dN_\bullet(t)$$

where $Y_\bullet(t) = Y_{\bullet 1}(t) + Y_{\bullet 2}(t)$ and $N_\bullet(t) = N_{\bullet 1}(t) + N_{\bullet 2}(t)$.

Show that under the null hypothesis of equal survival distributions in the two groups it holds that $\mathrm{E}(Z(\tau)) = 0$

You can use that the under the null hypothesis the counting processes $N_{\bullet j}(t)$ have cumulative intensity processes $\Lambda_j(t) = \int_0^t Y_{\bullet j}(t)\alpha(s)ds$ where $\alpha(s)$ is the common hazard function under the null hypothesis.

Below you see `R`-output from a log-rank test for the difference between groups where `surg=0` is the group with short time to surgery. Define the terms "Observed" and "Expected" in the table.

Comment on the degree of difference between the groups.

```
> survdiff(Surv(time,status)~surg)

         N Observed Expected (O-E)^2/E (O-E)^2/V
surg=0 682      316      337      1.32       5.2
surg=1 247      136      115      3.87       5.2

 Chisq= 5.2  on 1 degrees of freedom, p= 0.02
```

d) A next step in an analysis would often be to use a regression model to analyze the data such as the proportional hazards model which with one covariate can be stated as

$$\alpha(t|x_i) = \alpha_0(t)\exp(\beta x_i)$$

where $\alpha(t|x_i)$ is the hazard for an individual with one covariate x_i and $\alpha_0(t)$ a baseline hazard. The regression parameter β is often estimated by Cox-regression maximizing the partial likelihood

$$L(\beta) = \prod_{i=1}^n \prod_{0 \le t \le \tau} \left(\frac{\exp(\beta x_i)}{\sum_{j=1}^n Y_j(t)\exp(\beta x_j)} \right)^{dN_i(t)}$$

where $N_i(t)$ counts events and $Y_i(t)$ indicates at risk status (at time $t-$) for individual $i = 1, \ldots, n$.

On the next page you see `R`-output from a Cox-regression on the colon cancer data with the binary covariate x_i indicating that waiting time to surgery was long. Comment on the results.

In particular discuss what additional information this analysis gives compared with the log-rank test in question c). For this part you should find the score function $U(\beta)$ for estimating β evaluated in $\beta = 0$ and compare it to the log-rank statistic $Z(\tau)$.

(Continued on page 3.)

```
> summary(coxph(Surv(time,status)~surg))

       coef exp(coef) se(coef)     z Pr(>|z|)
surg 0.2333    1.2627   0.1026 2.274   0.0229 *
---

     exp(coef) exp(-coef) lower .95 upper .95
surg     1.263     0.7919     1.033     1.544

Likelihood ratio test= 5.01  on 1 df,   p=0.03
Wald test            = 5.17  on 1 df,   p=0.02
Score (logrank) test = 5.2  on 1 df,   p=0.02
```

Problem 2

In this problem we will consider the so-called accelerated failure time (AFT) model. This model can be specified by assuming that given a covariate x the distribution of a survival time T is given through the relation

$$\log(T) = \mu - \beta' x + \sigma W$$

where β is a regression parameter, W a specific random variable typically with mean zero and μ and σ location and dispersion parameters generating a family of distributions $\mu + \sigma W$.

a) Show that the survival function for T given x can be expressed as

$$S(t|x) = \mathrm{P}(T > t) = S_0(\exp(\beta' x)t)$$

where then $S_0(t) = \mathrm{P}(\exp(\mu + \sigma W) > t)$.

Give an explanation for the term AFT-model based on this representation.

b) Show that in general the hazard function of T given x can be written as $\alpha(t|x) = \exp(\beta' x)\alpha_0(\exp(\beta' x)t)$ where $\alpha_0(t) = \alpha(t|0) = -\frac{d}{dt}\log(S_0(t))$ is the hazard function when $x = 0$.

c) Assume now that $S_0(t) = \exp(-bt^k)$ for parameters $b > 0$ and $k > 0$, so that $\exp(\mu + \sigma W)$ has a Weibull distribution. Demonstrate that then the accelerated failure time model is also a proportional hazards model $\alpha(t|x_1) = \exp(\gamma' x)h_0(t)$ with a constant hazard ratio.

Determine the relation between the regression coefficient β in the accelerated failure time model and γ in the proportional hazards models.

(Continued on page 4.)

Problem 3

A special case of the additive hazards model is given by specifying the intensity processes for counting processes $N_i(t)$ for individuals $i = 1, \ldots, n$ as, with one scalar covariate x_i,

$$\lambda_i(t) = Y_i(t)(\beta_0(t) + \beta x_i) = Y_i(t)(\alpha_0(t) + \beta(x_i - \bar{x}(t)),$$

where β is a regression coefficient, $\beta_0(t)$ a baseline hazard (intercept function) for when $x_i = 0$ and the $Y_i(t)$ are the indicators that individuals $i = 1, \ldots, n$ are at risk at just before time t.

Furthermore, note the model reformulation by $\bar{x}(t) = \sum_{i=1}^n x_i Y_i(t) / \sum_{i=1} Y_i(t)$ and centered intercept function $\alpha_0(t) = \beta_0(t) + \beta\bar{x}(t)$.

Let also $\Lambda_i(t) = \int_0^t \lambda_i(s)ds$ be the cumulative intensities for the $N_i(t)$ and assume that we can observe events in the interval $[0, \tau]$. In the following you can use that $M_i(t) = N_i(t) - \Lambda_i(t)$ are uncorrelated (orthogonal) martingales with predictable variation processes $\langle M \rangle_i(t) = \Lambda_i(t)$ and optional variance processes $[M]_i(t) = N_i(t)$.

a) Show that

$$\int_0^\tau \sum_{i=1}^n (x_i - \bar{x}(t))dN_i(t) = \beta \int_0^\tau \sum_{i=1}^n Y_i(t)(x_i - \bar{x}(t))^2 dt + M_n(\tau)$$

where $M_n(t) = \int_0^t \sum_{i=1}^n (x_i - \bar{x}(s))dM_i(s)$ is a zero-mean martingale.

Explain why this suggest

$$\hat{\beta} = \frac{\int_0^\tau \sum_{i=1}^n (x_i - \bar{x}(t))dN_i(t)}{\int_0^\tau \sum_{i=1}^n Y_i(t)(x_i - \bar{x}(t))^2 dt}$$

as an estimator of β.

b) Assume that $A_n = \frac{1}{n}\int_0^\tau \sum_{i=1}^n Y_i(t)(x_i - \bar{x}(t))^2 dt \to a$ and that $B_n = \frac{1}{n}\int_0^\tau \sum_{i=1}^n (x_i - \bar{x}(t))^2 dN_i(t) \to b$ in probability.

Argue that then,

$$\sqrt{n}(\hat{\beta} - \beta) \to \mathrm{N}(0, b/a^2)$$

in distribtion as $n \to \infty$, so that $\hat{\beta}$ is approximately normal with expectation β and variance $b/(na^2)$.

c) Suggest an estimator for the integrated intercept (or cumulative baseline) function $B_0(t) = \int_0^t \beta_0(s)ds$.

Give a reason for your answer.

END