# 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 5 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.

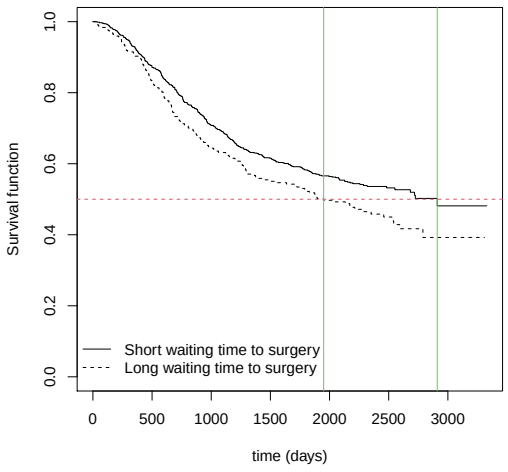
## Problem 1

a) By definition the hazard $\alpha(t) = f(t)/S(t) = -\frac{dS(t)}{S(t)}$ and so the cumulative hazard equals $\Lambda(t) = \int_0^t \lambda(s)ds = -\int_0^t \frac{dS(s)}{S(s)} = -\log(S(t))$ giving $S(t) = \exp(-\int_0^t \alpha(s)ds)$.

b) The Kaplan-Meier estimators are given as

$$\widehat{S}(t) = \prod_{s \le t} \left[1 - \frac{dN_{\bullet j}(s)}{Y_{\bullet j}(s)}\right]$$

We see that the the survival function for those with long waiting time is lower than for those with short waiting time, so mortality is highest in the long waiting time. For instance the median survival time is approximately 2000 (days) with long waiting and about 3000 days with short waiting.




*(Continued on page 2.)*

c) We can rewrite hazards for the groups, since under the null $\alpha_1(t) = \alpha_2(t) = \alpha(t)$,

$$\begin{aligned} Z(\tau) &= N_{\bullet 2}(\tau) - E_2(\tau) = N_{\bullet 2}(\tau) - \int_0^\tau \frac{Y_{\bullet 2}(t)}{Y_\bullet(t)} dN_\bullet(t) \\ &= \int_0^\tau \frac{Y_{\bullet 1}(t)}{Y_\bullet(t)} dN_{\bullet 2}(t) - \int_0^\tau \frac{Y_{\bullet 2}(t)}{Y_\bullet(t)} dN_{\bullet 1}(t) \\ &= (\int_0^\tau \frac{Y_{\bullet 1}(t)}{Y_\bullet(t)} d\Lambda_{\bullet 2}(t) - \int_0^\tau \frac{Y_{\bullet 2}(t)}{Y_\bullet(t)} d\Lambda_{\bullet 1}(t)) - (\int_0^\tau \frac{Y_{\bullet 1}(t)}{Y_\bullet(t)} dM_{\bullet 2}(t) - \int_0^\tau \frac{Y_{\bullet 2}(t)}{Y_\bullet(t)} dM_{\bullet 1}(t)) \\ &= (\int_0^\tau \frac{Y_{\bullet 1}(t)Y_{\bullet 2}(t)}{Y_\bullet(t)} \alpha(t)dt - \int_0^\tau \frac{Y_{\bullet 2}(t)Y_{\bullet 1}(t)}{Y_\bullet(t)} \alpha(t)dt) + M^*(\tau) = 0 + M^*(\tau) \end{aligned}$$

where $M^*(t) = -(\int_0^t \frac{Y_{\bullet 1}(s)}{Y_\bullet(s)} dM_{\bullet 2}(s) - \int_0^t \frac{Y_{\bullet 2}(s)}{Y_\bullet(s)} dM_{\bullet 1}(s))$ is a martingale with expectation zero. Thus $\mathrm{E}(Z(\tau) = 0$ under the null hypothesis.

We have $O_1 = N_{\bullet 1}(\tau) = 316, O_2 = N_{\bullet 2}(\tau) = 136, E_1 = \int_0^\tau \frac{Y_{\bullet 2}(t)}{Y_\bullet(t)} dN_{\bullet 1}(t) = 337$ and $\int_0^\tau \frac{Y_{\bullet 2}(t)}{Y_\bullet(t)} dN_\bullet(t) = 115$. Thus there are less events in the short wait group compared to the "expected" under null, and correspondingly more in the long wait group than "expected" and so higher mortality for long wait group in accordance with what observed from the Kaplan-Meier plots.

We also see that the difference is statistically significant at a standard 5 percent level with a p-value of 0.02.

```
> 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) The R-output shows a hazard ratio of $1.26 = \exp(\hat{\beta})$ between the long wait and short wait groups. The difference is significant since the confidence interval for $\exp(\beta)$ equals (1.03,1.54) and does not overlap with 1.

Also we see that the p-values lies in range 0.02-0.03 depending on which test is used. In particular the score test has a p-value of 0.02 for a test statistic of 5.2. This is the same value as the test statistic from the log-rank test. This is no coincidence since the value reported is identical to (assuming no ties in the data)

$$\frac{Z(\tau)^2}{\widehat{\mathrm{Var}}(Z(\tau))}$$

This is realized from considering

$$\log(L(\beta)) = \int_0^\tau \sum_{i=1}^n (\beta x_i - \log(\sum_{j=1}^n Y_j(t) \exp(\beta x_j)) dN_i(t)$$

*(Continued on page 3.)*

from which we obtain a score function

$$U(\beta) = \sum_{i=1}^{n} \int_0^\tau (x_i - \frac{\sum_{j=1}^n Y_j(t) x_j \exp(\beta x_j)}{\sum_{j=1}^n Y_j(t) \exp(\beta x_j)}) dN_i(t)$$

and evaluated in $\beta = 0$ we get exactly

$$U(0) = \int_0^\tau (x_i - \frac{\sum_{i=1}^n Y_i(t) x_i}{\sum_{i=1}^n Y_i(t)}) dN_i(t) = N_{\bullet 2}(\tau) - \int_0^\tau \frac{Y_{\bullet 2}(t)}{Y_\bullet(t)} dN_\bullet(t) = Z(\tau)$$

```
> 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

a) We have

$$\begin{aligned} S(t|x) &= \mathrm{P}(T > t) = \mathrm{P}(\log(T) > \log(t)) = \mathrm{P}(\mu - \beta' x + \sigma W > \log(t)) \\ &= \mathrm{P}(\mu + \sigma W > \log(t) + \beta' x) = \mathrm{P}(\exp(\mu + \sigma W) > t \exp(\beta' x)) \\ &= S_0(\exp(\beta' x) t) \end{aligned}$$

This survival function depend on time $t$ multiplied by a acceleration factor $\exp(\beta' x)$, so we can consider time as moving $\exp(\beta' x)$ faster with covarite $x$.

b) To get to the hazard function we use the representation from Problem 1, question a): $S(t) = \exp(-\int_0^t \alpha(s) ds)$ which correspond to

$$\alpha(t) = \frac{d}{dt}(-\log(S(t)) = \frac{-dS(t)}{S(t)}$$

Hence the hazard in an AFT model is given as

$$\alpha(t|x) = \frac{\exp(\beta' x)(-S_0'(\exp(\beta' x) t)}{S_0(\exp(\beta' x) t)} = \exp(\beta' x) \alpha_0(\exp(\beta' x) t)$$

*(Continued on page 4.)*

c) We can use the first representation $S(t|x) = S_0(\exp(\beta' x)t)$ which with $S_0(t) = \exp(-bt^k)$ gives

$$S(t|x) = \exp(-b(\exp(\beta' x)t)^k) = \exp(-b(\exp(k\beta' x)t^k) = \exp(-b(\exp(\gamma' x)t^k)$$

with is also the survival function of a Weibull distribution. The hazard corresponding to this survival function is given as

$$\alpha(t|x) = bk\exp(\gamma' x)t^{k-1} = \exp(\gamma' x)bkt^{k-1} = \exp(\gamma' x)h_0(t)$$

where $h_0(t) = bkt^{k-1}$ is the baseline corresponding to the Weibull survival function $S_0(t) = \exp(-bt^k)$. We recognize this model as a proportional hazards model with proportionality factor $\exp(\gamma' x) = \exp(k\beta' x)$, thus the regression parameter in the proportional hazard model equals $\gamma = k\beta$ compared to the model $\log(T) = \mu - \beta' x + \sigma W$.

## Problem 3

a) We have that $\int_0^\tau \sum_{i=1}^n (x_i - \bar{x}(t))dN_i(t)$

$= \int_0^\tau \sum_{i=1}^n (x_i - \bar{x}(t))(\lambda_i(t)dt + dM_i(t))$

$= \int_0^\tau \sum_{i=1}^n (x_i - \bar{x}(t))Y_i(t)(\alpha_0(t) + \beta(x_i - \bar{x}(t)))dt + dM_i(t))$

$= \int_0^\tau \alpha_0(t) \sum_{i=1}^n Y_i(t)(x_i - \bar{x}(t))dt + \beta \int_0^\tau \sum_{i=1}^n Y_i(t)(x_i - \bar{x}(t))(x_i - \bar{x}(t)))dt$

$\quad + \int_0^\tau \sum_{i=1}^n (x_i - \bar{x}(t))dM_i(t))$

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

where $M_n(t) = \int_0^t \sum_{i=1}^n (x_i - \bar{x}(s))dM_i(s)$ is a zero-mean martingale. The term with $\alpha_0(t)$ vanish since $\sum_{i=1}^n Y_i(t)(x_i - \bar{x}(t)) = 0$.

We can then write

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

where the nominator in the second term is a martingale and so has expectation zero. We thus have $\hat{\beta} = \beta +$ a term that must have mean close to zero. This is sufficient for establishing a meaningful estimator of $\beta$, but a fuller derivation requires that the second term also converge to zero as $n \to \infty$. This involves consideration of the variance of $M_n(\tau)$ in relation to the nominator $\int_0^\tau \sum_{i=1}^n Y_i(t)(x_i - \bar{x}(t))^2 dt$ - as in question b).

b) From a) we have that

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

*(Continued on page 5.)*

so the result will follow from $\frac{1}{\sqrt{n}}M_n(\tau) \to N(0,b)$ since $A_n \to a$ by assumption.

But by the martingale central limit theorem it follows under natural assumptions that $\frac{1}{\sqrt{n}}M_n(t)$ converges to a normal distribution where the limit variance is obtained considering (predictable and/or optional) variance processes for $\frac{1}{\sqrt{n}}M_n(t)$

Specifically the variance of $M_n(t)$ is given as $\mathrm{E}(\langle M_n\rangle(t)$ where $\langle M_n\rangle(t)$ is the predictable variance process of $M(t)$. This can be expressed as

$$\begin{aligned}\langle M_n\rangle(t) &= \langle \textstyle\int_0^t \sum_{i=1}^n (x_i - \bar{x}(s))dM_i(s)\rangle = \int_0^t \sum_{i=1}^n (x_i-\bar{x}(s))^2 d\langle M_i(s)\rangle \\ &= \textstyle\int_0^t \sum_{i=1}^n (x_i-\bar{x}(s))^2\lambda_i(s)ds\end{aligned}$$

and can be estimated by

$$nB_n = \int_0^t \sum_{i=1}^n (x_i-\bar{x}(s))^2 Y_i(s)dN_i(s)$$

which incidentally equal the optional variation process $[M](t)$ for $M(t)$.

Assume that $A_n = \frac{1}{n}\int_0^\tau \sum_{i=1}^n Y_i(t)(x_i-\bar{x}(t))^2dt \to a$ 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 $\beta$ and variance $b/(na^2)$.

c) The Nelson-Aalen estimator $\int_0^t \frac{dN(s)}{Y(s)}$ where $N(t) = \sum_{i=1}^n N_i(t)$ and $Y_i(t) = \sum_{i=1}^n Y_i(s)$ under this additive hazards model can be written as

$$\begin{aligned}\textstyle\int_0^t \frac{dN(s)}{Y(s)} &= \textstyle\int_0^t \frac{\sum_{i=1}^n Y_i(s)(\beta_0(s)+\beta x_i)ds}{Y(s)} + \int_0^t \frac{\sum_{i=1}^n dM_i(s)}{Y(s)} \\ &= \textstyle\int_0^t J(s)\beta_0(s)ds + \beta\int_0^t \frac{\sum_{i=1}^n Y_i(s)x_i}{Y(s)}ds + M^*(t) \\ &= B^*(t) + \beta\textstyle\int_0^t \bar{x}(s)ds + M^*(t)\end{aligned}$$

where $J(s) = I(Y(s)>0)$, $B^*(t) = \int_0^t J(s)\beta_0(s)ds$ and $M^*(t)$ is a martingale with expectation zero. This suggests

$$\hat{B}_0(t) = \int_0^t \frac{dN(s)}{Y(s)} - \hat{\beta}\int_0^t \bar{x}(s)ds$$

as an estimator of $B_0(t)$.

END