Department of Mathematical Sciences

Examination paper for TMA4275 Lifetime analysis

Examination date: 02.06.2023

Examination time (from-to): 09.00 - 13.00

Permitted examination support material: C

- Approved calculator
- Tabeller og formler i statistikk, Akademika
- A yellow sheet of paper (A5 with a stamp) with personal handwritten formulas and notes

Academic contact during examination: Håkon Tjelmeland
Phone: 4822 1896

Academic contact present at the exam location: Yes, approximately 11.00

OTHER INFORMATION

Get an overview of the question set before you start answering the questions.

Language: Your answers may be given in Norwegian or English

Read the questions carefully and make your own assumptions. If a question is unclear/vague, make your own assumptions and specify them in your answer. Only contact academic contact in case of errors or insufficiencies in the question set. Address an invigilator if you wish to contact the academic contact. Write down the question in advance.

InsperaScan: For all questions you are meant to answer on handwritten sheets. At the bottom of the question you will find a seven-digit code. Fill in this code in the top left corner of the sheets you wish to submit. We recommend that you do this during the exam. If you require access to the codes after the examination time ends, click “Show submission”.

Weighting: The exam consists of seven parts (1, 2, 3, 4a, 4b, 5a and 5b). These seven parts are given equal weight in the evaluation of your solution.

Notifications: If there is a need to send a message to the candidates during the exam (e.g. if there is an error in the question set), this will be done by sending a notification in Inspera. A dialogue box will appear. You can re-read the notification by clicking the bell icon in the top right-hand corner of the screen.

Withdrawing from the exam: If you become ill or wish to submit a blank test/withdraw from the exam for another reason, go to the menu in the top right-hand corner and click “Submit blank”. This cannot be undone, even if the test is still open.

Access to your answers: After the exam, you can find your answers in the archive in Inspera. Be aware that it may take a working day until any hand-written material is available in the archive.

1 Kaplan-Meyer estimator

Introduction: The following table contains some observed lifetimes and some right censored times. The censoring times are marked with an asterix ($\star$).

$$0.567^{\star}, 0.65^{\star}, 0.70, 0.77^{\star}, 1.04, 1.13^{\star}, 1.15, 1.63^{\star}$$

Problem: Use the above data to compute by hand the Kalman-Meier estimate for the survival function. Include in your solution enough of intermediate calculations so that is is completely clear how you have found your results. For evaluations of sums, products etc you may of course use your calculator.

Make a plot of the resulting estimated survival function for $t \in [0, 1.75]$. Use the plot to find an estimate for the median survival time.

You should preferably answer this problem by writing by hand on a sheet of paper.

Format

Words: 0

Maximum marks: 10

2 Hazard rate and survival function

Introduction: Let T be a lifetime for which the density function is given by

$$f(t) = \varphi e^{-\theta t} \exp\left\{-\tfrac{\varphi}{\theta}\left(1 - e^{-\theta t}\right)\right\} \text{ for } t \geq 0,$$

where $\theta > 0$ and $\varphi > 0$ are parameters.

Problem: Derive formulas for the hazard rate of T, $\alpha(t)$, and for the survival function $S(t)$.

You should preferably answer this problem by writing by hand on a sheet of paper.

Format B I U x_2 x^2 I_x Ω Σ

Words: 0

Maximum marks: 10

3 Estimator for the variance of the Nelson-Aalen estimator

Introduction: Let $N(t)$ be a counting process with a multiplicative intensity process $\lambda(t) = Y(t)\alpha(t)$, where $\alpha(t)$ is a hazard rate and $Y(t)$ is a predictable process. As usual we let $A(t) = \int_0^t \alpha(s)ds$ denote the corresponding integrated hazard rate. As discussed in class it is common to estimate $A(t)$ by the Nelson-Aalen estimator

$$\widehat{A}(t) = \int_0^t \frac{J(s)}{Y(s)} dN(s),$$

where $J(s) = I(Y(s) > 0)$ equals one if $Y(s) > 0$ and zero otherwise. Moreover, we have discussed in class that the common estimator for the variance of the Nelson-Aalen estimator is

$$\hat{\sigma}^2(t) = \sum_{j:T_j \leq t} \frac{1}{Y(T_j)^2},$$

where the sum is over all observed event times that are less or equal to t. Just as we have done in class, we in the following define

$$A^\star(t) = \int_0^t J(s)\alpha(s)ds.$$

In the following you can use without proof that the Doob-Meyer decomposition of a counting process is

$$N(t) = \int_0^t \lambda(s)ds + M(t),$$

where $\lambda(t)$ is the intensity process of $N(t)$ and $M(t)$ is a mean zero martingale.

Problem: Use the Doob-Meyer decomposition of $N(t)$ to show that

$$\widehat{A}(t) - A^\star(t) = \int_0^t \frac{J(s)}{Y(s)} dM(s).$$

Explain why this implies that $\widehat{A}(t) - A^\star(t)$ is a mean zero martingale.

Find an expression for the optional variation process of $\widehat{A}(t) - A^\star(t)$ and use this to show that

$$\left[\widehat{A} - A^\star\right](t) = \hat{\sigma}^2(t).$$

Explain why this result implies that $\hat{\sigma}^2(t)$ is an unbiased estimator of the variance of $\widehat{A}(t) - A^\star(t)$.

You should preferably answer this problem by writing by hand on a sheet of paper.

Format

Words: 0

Maximum marks: 10

4 Cox model and partial likelihood

Introduction: Assume we have n machine components of a certain type. The components are from two different manufacturers and the various components are operating in different environments (temperature, pressure etc). We denote the two manufacturers by manufacturer A og manufacturer B, and set $x_{i1} = 0$ if component number i is from manufacturer A and $x_{i1} = 1$ if it is from manufacturer B. We measure the effect of the environment by a (time invariant) *stress factor*, and the value of this stress factor for component number i we denote by $x_{i2} \in \mathbb{R}$.

We assume component i to fail according to an intensity process

$$\lambda_i(t) = \alpha_0(t) \exp\{\beta_1 x_{i1} + \beta_2 x_{i2}\},$$

where $\alpha_0(t)$ is an unspecified baseline hazard function which we assume to be the same for all components, β_1 and β_2 are parameters that we want to estimate, and t is time from the component was taken into use. We assume the components to fail independently of each other. Moreover, when a component fails we assume that it is immediately repaired and that after the repair the intensity for a new failure is again given by the $\lambda_i(t)$ specified above. Thus, all n components are under risk of failure at all times, and the intensity of a failure to occur is not influenced by the number of previous failures for the same component.

Assume we observe each of the n components from time zero to some time τ, and we assume no censoring before time τ. As usual we let T_j denote the jth observed failure time and assume that this failure occurred for component number i_j.

Problems:

a) Starting from the formula for the partial likelihood for a general Cox model that we found in class,

$$L(\beta) = \prod_j \frac{\exp\{\beta^T x_{i_j}(T_j)\}}{\sum_{\ell \in \mathcal{R}_j} \exp\{\beta^T x_\ell(T_j)\}},$$

where $\mathcal{R}_j$ is the set of components under risk at time T_j, derive a formula for the log-partial likelihood function in the situation defined above, and show that it can be expressed as

$$\ell(\beta) = \beta_1 m^{(1)} + \beta_2 \sum_j x_{i_j 2} - m \ln\left(h_0(\beta_2) + e^{\beta_1} h_1(\beta_2)\right),$$

where $m^{(1)}$ is the number of failures occurred for components from manufacturer B, m is the total number of failures occurred for the n components, and

$$h_0(\beta_2) = \sum_{\ell : x_{\ell 1} = 0} \exp\{\beta_2 x_{\ell 2}\} \quad \text{and} \quad h_1(\beta_2) = \sum_{l : x_{\ell 1} = 1} \exp\{\beta_2 x_{\ell 2}\}.$$

b) If possible, find explicit formulas for the maximum partial likelihood estimators for β_1 and β_2. If this is not possible, if possible optimise analytically with respect to one of β_1 and β_2 and find an analytical expression for the profile log-partial likelihood for the other parameter. If it is not possible to optimise analytically with respect to any of the two parameters, explain why.

You should preferably answer these problems by writing by hand on a sheet of paper.

Format

Words: 0

Maximum marks: 30

5 Estimation in a parametric model

Introduction: The R output below shows some results when estimating a (parametric) Weibull regression model to a survival data set of 241 individuals with monoclonal gammopathy of undetermined significance (MGUS). The fitted model has two covariates: age (age in years at detection of MGUS) and sex (male or female). The covariate age is continuous, whereas the covariate sex is binary. In the data set the values of the covariate age varies from 34 to 90. From detection of MGUS, each individual is followed up until a time denoted *futime* (specified in number of days) in the dataset. In the data set the variable *death* is an indicator that equals one if the follow up was until death, and zero otherwise.

R commands and R output:

```
> result = survreg(Surv(futime,death) ~ age + sex,data=mgus,dist="weibull")
> summary(result)

Call:
survreg(formula = Surv(futime, death) ~ age + sex, data = mgus, dist = "weibull")
              Value      Std. Error   z        p
(Intercept)  11.24975    0.29247   38.46   < 2e-16
age          -0.03995    0.00452   -8.83   < 2e-16
sexmale      -0.16387    0.09744   -1.68    0.093
Log(scale)   -0.33389    0.05650   -5.91    3.4e-09

Scale= 0.716

Weibull distribution
Loglik(model)= -2130.3   Loglik(intercept only)= -2170.9
Chisq= 81.18 on 2 degrees of freedom, p= 2.4e-18
Number of Newton-Raphson Iterations: 7
n= 241

> result$var

               [,1]            [,2]            [,3]            [,4]
[1,]  0.085537586   -1.279591e-03   -2.805729e-03    2.141440e-03
[2,] -0.001279591    2.045819e-05   -4.167446e-05   -4.039657e-05
[3,] -0.002805729   -4.167446e-05    9.495064e-03   -2.687217e-04
[4,]  0.002141440   -4.039657e-05   -2.687217e-04    3.192647e-03
```

As we have discussed in class, the parameterisation used by the R function *survreg* differs from the parameterisation used in our textbook Aalen et al. (2008). Denoting the parameters used in the R function *survreg* by μ (Intercept), γ_1 (age), γ_2 (sexmale) and τ (Log(scale)) the parameterisation of the intensity process for individual i in *survreg* is

$$\lambda_i(t) = \exp\left\{-(\tau + \mu e^{-\tau}) \cdot t^{\exp\{-\tau\}-1} \cdot \exp\{-\gamma_1 e^{-\tau} \cdot \mathrm{age}_i - \gamma_2 e^{-\tau} \cdot \mathrm{sexmale}_i\right\},$$

where age_i and $\mathrm{sexmale}_i$ denote the age at detection of MGUS (in years) and sex of individual number i, respectively. Using the parameterisation in Aalen et al. (2008) the hazard rate for the same individual reads

$$\lambda_i(t) = bt^{k-1} \exp\{\beta_1 \cdot \mathbf{age}_i + \beta_2 \cdot \mathbf{sexmale}_i\}.$$

The relation between the two parameterisations can thereby be expressed as

$$b = \exp\{-[\tau + \mu e^{-\tau}]\}, \quad \beta_1 = -\gamma_1 e^{-\tau}, \quad \beta_2 = -\gamma_2 e^{-\tau} \quad \text{and} \quad k = e^{-\tau}.$$

Problems:

a) Find the estimated intensity process for a female who was 45 years when MGUS was detected.

Find a 95% confidence interval for the parameter k.

Based on the results above, discuss whether you would expect an exponential regression model to give a good fit to the data set.

b) For a male who was 50 years when MGUS was detected, find formula for the estimated survival function, and in particular find the estimated probability that such an individual will survive for $10\,000$ days.

Discuss how you for such an individual also can find a 95% confidence interval for $S(10\,000) = P(T > 10\,000)$. You do NOT need to do all the calculations, but for the specific situation in question you should formulate mathematically how you would derive such an interval.

You should preferably answer these problems by writing by hand on a sheet of paper.

Format

Words: 0

Maximum marks: 20