NTNU
Norwegian University of
Science and Technology

Department of Mathematical Sciences

Examination paper for **TMA4275 Lifetime analysis**

Academic contact during examination: Bo Lindqvist

Phone: 97 58 94 18

Examination date: June 10, 2017

Examination time (from–to): 09:00–13:00

Permitted examination support material: *Tabeller og formler i statistikk*, Tapir Forlag, K. Rottmann: *Matematisk formelsamling*, Calculator Casio fx-82ES PLUS, CITIZEN SR-270X, CITIZEN SR-270X College or HP30S, one yellow A4-sheet with your own handwritten notes.

Other information:
Note that you should explain your reasoning behind your answers. You may write in English and/or Norwegian. You may write with a pencil.

Language: English

Number of pages: 9

Number of pages enclosed: 0

Checked by:

Informasjon om trykking av eksamensoppgave	
Originalen er:	
1-sidig ☐	**2-sidig** ☒
sort/hvit ☒	**farger** ☐
skal ha flervalgskjema ☐	

Date Signature

Problem 1

A store selling a particular mobile phone model give their customers a one year (365 days) guarantee and offer to repair any phones that have a failure before the guarantee time expires. In such cases the failure time T_i is recorded as the variable y_i in Table 1 along with a censoring indicator $\delta_i = 1$. In cases where no failure occurred before the end of the guarantee time, y_i is assigned a value of 365 and the censoring indicator $\delta_i = 0$. Also recorded is the sex of each customer and the average number of minutes per day the customer used the phone (such usage data are reported over the mobile network back to phone manufacturer by this particular phone model).

i	y_i	δ_i	sex	usage
1	16	1	female	6
2	18	1	female	18
3	20	1	female	7
4	32	1	female	12
5	56	1	female	22
6	115	1	female	1
7	116	1	female	0
8	251	1	female	8
9	253	1	female	13
10	281	1	female	6
11	303	1	female	3
12	365	0	female	4
13	365	0	female	1
14	1	1	male	70
15	5	1	male	30
16	44	1	male	26
17	59	1	male	24
18	70	1	male	16
19	100	1	male	14
20	161	1	male	13

Table 1: Mobile phone data of problem 1

a) Fig. 1 shows the Kaplan-Meier estimate $\hat{R}(t)$ of the reliability function $R(t) = P(T > t)$ of phones used by male and female customers. Calculate $\hat{R}(100)$ for males by hand. Based on the estimates of $R(t)$, compute estimates of the median lifetime of phones used by males and females. Also explain how the expected lifetime of phones of male and female users can or cannot be

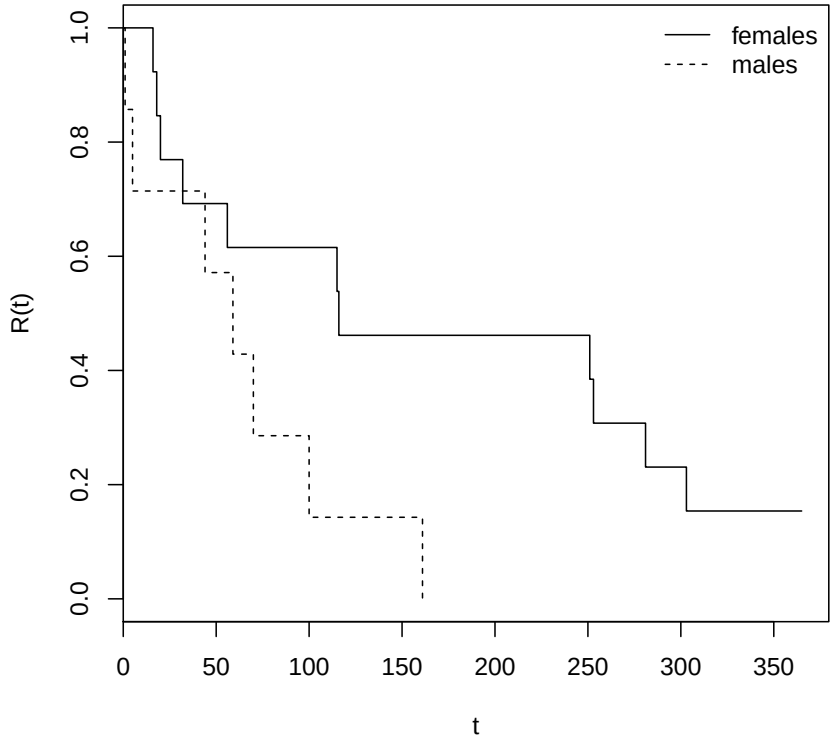

Figure 1: Kaplan Meier estimates of $R(t)$ for males and females based on the data given in Table 1

computed based on the plot.

b) The following output from R shows the result of a log-rank test comparing groups defined by the variable `sex`.

```
> survdiff(Surv(y, delta) ~ sex)
Call:
survdiff(formula = Surv(y, delta) ~ sex)

            N Observed Expected (O-E)^2/E (O-E)^2/V
sex=female 13       11    14.27     0.748      4.04
sex=male    7        7     3.73     2.858      4.04

 Chisq= 4  on 1 degrees of freedom, p= 0.0445
```

Briefly state the null and alternative hypothesis being tested. What is the observed and expected number of failures under H_0 in each group at the time of the failure recorded for unit $i = 3$ that failed at $y_i = 20$? How do these numbers relate to the numbers in columns `Observed` and `Expected` in the above R output?

Based on the test, can we conclude that the sex of the user has a direct causal effect on the lifespan of this phone model?

c) We next fit three different Cox proportional hazards models as follows.

```
> cox1 <- coxph(Surv(y, delta) ~ sex)
> cox1
Call:
coxph(formula = Surv(y, delta) ~ sex)

        coef exp(coef) se(coef)    z     p
sexmale 1.06      2.89     0.55 1.93 0.054

Likelihood ratio test=3.59  on 1 df, p=0.0581
n= 20, number of events= 18
> logLik(cox1)
'log Lik.' -39.84702 (df=1)
> cox2 <- coxph(Surv(y, delta) ~ usage)
> cox2
Call:
coxph(formula = Surv(y, delta) ~ usage)

        coef exp(coef) se(coef)    z      p
usage 0.1262    1.1345   0.0397 3.18 0.0015
```

```
Likelihood ratio test=16.3  on 1 df, p=5.3e-05
n= 20, number of events= 18
> logLik(cox2)
'log Lik.' -33.47406 (df=1)
> cox12 <- coxph(Surv(y, delta) ~ sex + usage)
> cox12
Call:
coxph(formula = Surv(y, delta) ~ sex + usage)

           coef exp(coef) se(coef)     z      p
sexmale -0.3232    0.7238   0.6865 -0.47 0.6378
usage    0.1377    1.1477   0.0472  2.92 0.0035

Likelihood ratio test=16.6  on 2 df, p=0.000254
n= 20, number of events= 18
> logLik(cox12)
'log Lik.' -33.36336 (df=2)
```

Write the last model in mathematical notation and state its assumption. Based on likelihood ratio tests, is the effect of `sex` significant when `usage` is included in the model? Which model do you prefer out of the three alternatives? Based on your preferred model, by how much does the hazard change per minute increase in average daily phone usage?

d) For models with a single covariate, the Schoenfeld residual at the ith failure can be written as

$$\text{res}_i = x_i - \frac{\sum_{k \in R_i} x_k e^{\hat{\beta} x_k}}{\sum_{k \in R_i} e^{\hat{\beta} x_k}}, \tag{1}$$

where x_i is the value of the covariate of the unit that failed at the ith failure, R_i is the set of units at risk immediately before the ith failure, and $\hat{\beta}$ is the value of β maximizing the Cox partial likelihood. Give an interpretation of the terms in the above formula and explain how these residuals behave if the proportional hazard assumption holds. Also explain how these residuals would behave if the proportional effect of daily phone usage on the hazard function $z(t)$ is not constant over time but instead increasing.

Fig. 2 shows these Schoenfeld residuals for the `usage` covariate for model `cox2` plotted against the observed failure times. Judging the distribution by eye, do you see any strong evidence that the proportional hazard assumption is violated?

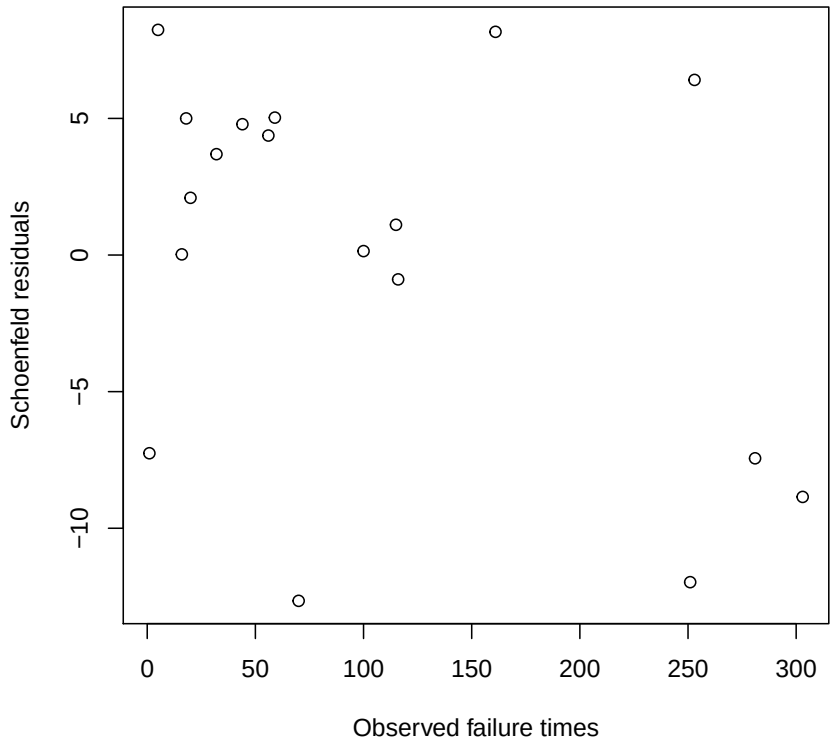

Figure 2: Schoenfeld residuals for the `usage` covariate for model `cox2` plotted against the observed failure times

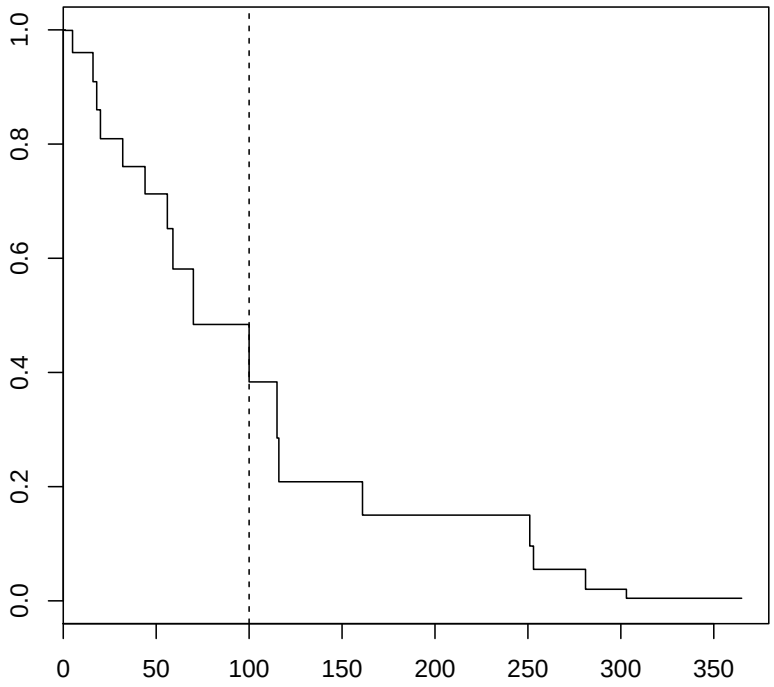

Figure 3: An estimate of the baseline survival function $R_0(t)$ under model `cox2` in point c).

e) Fig. 3 shows an estimate of the baseline survival function $R_0(t)$ under model `cox2` fitted above. Explain how $R_0(t)$ relates to the baseline hazard function $z_0(t)$ and how the survival function of a subject with covariate vector $\mathbf{x}_i$, $R(t; \mathbf{x}_i)$, relates to $R_0(t)$.

Based on Fig. 3, for a customer using the phone on average 10 minutes per day, find an estimate of the probability that the phone is still functioning after 100 days.

f) We next fit a parametric survival regression model assuming that the lifetimes follow a log-normal distribution.

```
> srmod <- survreg(Surv(y, delta) ~ usage, dist = "lognormal")
> summary(srmod)

Call:
survreg(formula = Surv(y, delta) ~ usage, dist = "lognormal")
              Value Std. Error      z        p
(Intercept)  5.4003     0.3245 16.642 3.43e-62
usage       -0.0800     0.0151 -5.304 1.13e-07
Log(scale)   0.0226     0.1696  0.133 8.94e-01

Scale= 1.02

Log Normal distribution
Loglik(model)= -99.3   Loglik(intercept only)= -108.1
Chisq= 17.5 on 1 degrees of freedom, p= 2.9e-05
Number of Newton-Raphson Iterations: 6
n= 20
```

Write down the model in mathematical notation and state its assumptions.

Using this new parametric model, suppose we want to re-estimate the same probability as in point e), that is, the probability $p = P(T > t)$, $t = 100$ days, for a user with daily average phone usage of 10 minutes. Express p in terms of the parameters of the survival regression model (say β_0, β_1 and $\ln \sigma$). Compute a numerical estimate $\hat{p}$ of p.

Also derive the necessary formulas for computing the approximate variance of $\hat{p}$ given the following variance-covariance matrix of the parameter estimates of the survival regression. You may express these in terms of the probability density function $\phi(z)$ and cumulative density function $\Phi(z)$ of the standard normal distribution. You do not need to carry out the final tedious numerical calculations.

```
> vcov(srmod)
             (Intercept)         usage    Log(scale)
(Intercept)  0.105292325 -0.0034234792  0.0041043409
usage       -0.003423479  0.0002275326 -0.0001212433
Log(scale)   0.004104341 -0.0001212433  0.0287608330
```

g) Find the expected lifetime (ET = MTTF) of the phone of a user with daily average phone usage of 10 minutes, again based on the model in point f).

In addition, assuming instead that average daily phone usage varies between different individuals according to an exponential distribution with expected value equal to 10 minutes, find the expected lifetime of the phone of a randomly chosen individual from the population.

Hint: A normally distributed variable $U \sim N(\mu, \sigma^2)$ has moment-generating function $M_U(t) = Ee^{tU} = e^{\mu t + \sigma^2 t^2/2}$ and an exponentially distributed variable X with mean θ has moment-generating function $M_X(t) = Ee^{tX} = 1/(1-\theta t)$.

Problem 2

In this problem we are studying the reliability of three computer servers $j = 1, 2, 3$ and record the times (in days) s_{ij}, $i = 1, 2, \ldots, n_j$ at which each server needed to be rebooted after initial installation during τ_j days of operation (Table 2).

a) Compute the total time on test $Y_1, Y_2, \ldots, Y_{10}$ at each ordered failure time.

Let H_0 denote the null hypothesis that there is no change in the failure rate with time and no difference between the computers in their failure rates. How are $Y_1, Y_2, \ldots$ distributed under the H_0? How are are $Y_1, Y_2, \ldots, Y_{10}$ distributed conditional on 10 failures occurring in total, again under H_0?

Draw a total-time-on-test plot of the data. Does the plot indicate that the failure rate is constant, increasing or decreasing with time?

What is the approximate distribution of $\sum_{j=1}^{10} Y_i$ under H_0? Carry out the Laplace test (similar to the Proschan-Barlow test for lifetime data) of H_0 using a significance level of $\alpha = 0.05$.

b) Assume that each failure s_{ij}, $i = 1, 2 \ldots, n_j$ for each computer $j = 1, 2, 3$ occur according to a non-homogeneous Poisson processes with intensity $w(t)$ on the intervals $(0, \tau_j)$. Write down the general form of the likelihood under this model.

j	τ_j	n_j	s_{ij}
1	50	1	21
1	100	2	75, 92
1	200	7	55, 122, 125, 173, 178, 190, 195

Table 2: Computer failure time data in problem 2

Suppose that the intensity is given by $w(t) = e^{\beta_0+\beta_1 t}$ where β_0 and β_1 are unknown parameters. Derive an expression for the log-likelihood function.

c) Using numerical methods, we maximize the log likelihood function in point b) and obtain maximum likelihood estimates and standard errors based on the observed Fisher information matrix equal to $\hat{\beta}_0 = -5.52 \pm 0.622$ and $\hat{\beta}_1 = 0.018 \pm 0.00435$. The maximum log likelihood is -43.17.

Carry out a Wald test (also referred to as a Z-test) and an approximate likelihood ratio test of the same null hypothesis H_0 as in point a) using a significance level of $\alpha = 0.05$.