BIOSTAT III: Survival Analysis for Epidemiologists in R: Take-home examination (answers)

Mark Clements

8–17 November, 2021

Instructions

- *If you find any issues with these answers, please inform Mark Clements using mark.clements@ki.se.*
- **The examination is individual-based: you are not allowed to cooperate with anyone**, although you are encouraged to consult the available literature. The examiner will use Urkund in order to assess potential plagiarism.
- The examination will be made available by noon on Wednesday 17 November 2021 and **the examination is due by 17:00 on Wednesday 24 November 2021**.
- The examination will be graded and results returned to you by Wednesday 1 December 2021.
- The examination is in two parts. To pass the examination, you need to score at least 7/12 for Part 1 focused on rates and general regression modelling and 13/25 for Part 2 on survival analysis.
- Do not write answers by hand: please use Word, LaTeX, Markdown or a similar format for your examination report and submit the report **as a PDF file**.
- Motivate all answers in your examination report. Define any notation that you use for equations. The examination report should be written in English.
- Email the examination report containing the answers **as a PDF file** to gunilla.nilsson.roos@ki.se **Write your name in the email, but do NOT write your name or otherwise reveal your identity in the document containing the answers.**

Part 1

We load the `blcaIT` dataset from the `Epi` package and show its documentation:

```
library(Epi) # blcaIT
data(blcaIT)
help("blcaIT",help_type="text")
```

```
Bladder cancer mortality in Italian males

Description:

     Number of deaths from bladder cancer and person-years in the
     Italian male population 1955-1979, in ages 25-79.
```

```
Format:

     A data frame with 55 observations on the following 4 variables:

          'age':  Age at death. Left endpoint of age class
        'period':  Period of death. Left endpoint of period
             'D':  Number of deaths
             'Y':  Number of person-years.
```

Note that age and period are in five-year intervals.

Q1

(a) The age-specific mortality rates by calendar period are shown in Figure 1. Carefully describe the pattern of rates by age and calendar period. (2 pts)

```
library(ggplot2) # ggplot
ggplot(transform(blcaIT, R=D/Y*1e5), aes(x=age,y=R,col=factor(period))) +
    geom_line() + xlab("Age (years)") + ylab("Mortality rate per 100,000")
```

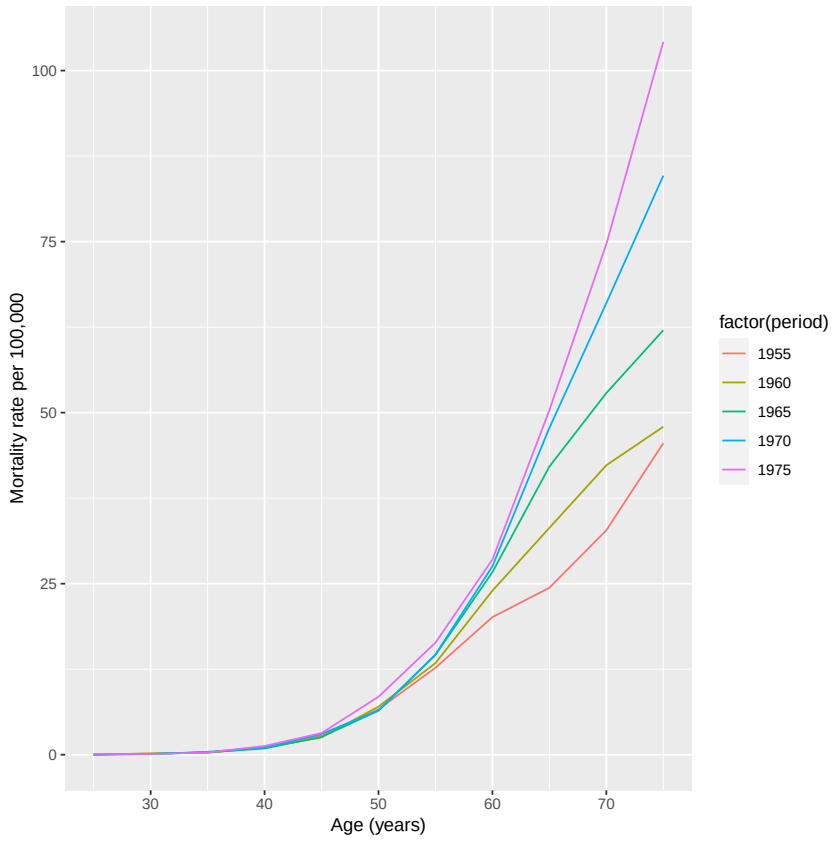

Figure 1: Age-specific bladder cancer mortality rates by calendar period, males, Italy, 1955-1979

Answer Mortality rates are very low prior to age 35 years, with a rapid increase in rates at older ages. At age 60-64 years, the rates across the periods 1955-1979 were approximately 25 per 100,000, while the rates at ages 75-79 years varied between less than 50 per 100,000 to higher than 100 per 100,000. There was strong evidence for an increase in rates over

the calendar periods, with the period 1975-1979 being higher than all of the other calendar periods for all age groups 50 years and over, and where those who died in 1955-1959 had lower rates from age 55 years and over. There was some evidence for non-proportionality by calendar period and age group, where the relative differences between the calendar periods tended to be greater in the older age groups.

(b) We fit a Poisson regression model for bladder cancer mortality in 1955-59 and 1975-1979, with main effects for age and calendar period. Write a formula for this regression model. As a reminder, please define your notation. (2 pts)

```
## create an indicator for the 1975-1979 calendar period
blcaIT2 = transform(blcaIT, period_1975 = ifelse(period==1975, 1, 0))
summary(glm(D ~ factor(age)+period_1975+offset(log(Y)), data=blcaIT2,
    subset=(period %in% c(1955,1975)), family=poisson))


Call:
glm(formula = D ~ factor(age) + period_1975 + offset(log(Y)),
    family = poisson, data = blcaIT2, subset = (period %in% c(1955,
        1975)))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.6214  -3.2045  -0.1907   2.9761   5.3182

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -14.43686    0.25842 -55.866  < 2e-16 ***
factor(age)30   0.56694    0.32913   1.723    0.085 .
factor(age)35   1.41980    0.29187   4.865 1.15e-06 ***
factor(age)40   2.71850    0.26800  10.144  < 2e-16 ***
factor(age)45   3.68092    0.26190  14.054  < 2e-16 ***
factor(age)50   4.59558    0.25983  17.687  < 2e-16 ***
factor(age)55   5.26136    0.25932  20.289  < 2e-16 ***
factor(age)60   5.76261    0.25892  22.257  < 2e-16 ***
factor(age)65   6.22596    0.25870  24.067  < 2e-16 ***
factor(age)70   6.59445    0.25868  25.493  < 2e-16 ***
factor(age)75   6.92308    0.25876  26.755  < 2e-16 ***
period_1975     0.58342    0.01662  35.105  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 37470.06  on 21  degrees of freedom
Residual deviance:   254.35  on 10  degrees of freedom
AIC: 438.94

Number of Fisher Scoring iterations: 4
```

Answer One representation of the regression equation is

$$E(D) = \exp\left(\beta_0 + \beta_1 I(30 \leq age < 35) + \beta_2 I(35 \leq age < 40) + \right.$$
$$\beta_3 I(40 \leq age < 45) + \beta_4 I(45 \leq age < 50) + \beta_5 I(50 \leq age < 55) + \beta_6 I(55 \leq age < 60) +$$
$$\beta_7 I(60 \leq age < 65) + \beta_8 I(65 \leq age < 70) + \beta_9 I(70 \leq age < 75) + \beta_{10} I(75 \leq age < 80) +$$
$$\left. \beta_{11} I(1975 \leq period < 1980) + \log(Y)\right)$$

where $E(D)$ is the expected count, β_0 is the log rate for those aged 25-29 years who died in the 1955-1959 calendar period, $\beta_2, \ldots, \beta_{10}$ are the log rate ratios comparing those aged 30-34 years, ..., 75-79 years compared with those aged 25-29 years, β_{11} is the log rate ratio comparing those diagnosed in the 1975-1979 period compared with those who died in the 1955-1959 calendar period, and Y is the person-time (so that $\log(Y)$ is the offset). We have used the $I()$ function as an indicator with a logical predicate, where the function is 1 if the predicate is true and 0 if the predicate is false. As a comment: the question related to the *fitted* model, rather than a general main effects model (where we could have, for example, fitted age as a linear effect).

(c) Based on the previous regression output, what is the mortality rate ratio and 95% confidence interval comparing the 1975-1979 calendar period with the 1955-1959 calendar period after adjusting for age? (2 pts)

Answer The rate ratio is exp(0.58342) = `1.792`, with 95% confidence interval (exp(0.58342-1.96*0.01662), exp(0.58342+1.96*0.01662)) = (`1.735`, `1.851`).

(d) To investigate how the mortality rate ratio varies by age, we now fit a model with an interaction between calendar period and age groups. Provide a formula for the regression model. (2 pts)

```
blcaIT3 = transform(blcaIT2,
    age_30_49 = (age>=30 & age < 50), # indicator for ages 30-49 years
    age_65_79 = (age>=65 & age < 80)) # indicator for ages 65-79 years
summary(glm(D ~ factor(age)+period_1975+period_1975:age_30_49+
period_1975:age_65_79+offset(log(Y)), data=blcaIT3,
    subset=(period %in% c(1955,1975)), family=poisson))


Call:
glm(formula = D ~ factor(age) + period_1975 + period_1975:age_30_49 +
    period_1975:age_65_79 + offset(log(Y)), family = poisson,
    data = blcaIT3, subset = (period %in% c(1955, 1975)))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4972  -0.6842  -0.0017   0.6461   1.5367

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -14.26361    0.25869 -55.139  < 2e-16 ***
factor(age)30              0.67175    0.33153   2.026  0.04275 *
factor(age)35              1.55714    0.29522   5.275 1.33e-07 ***
factor(age)40              2.85047    0.27152  10.498  < 2e-16 ***
factor(age)45              3.80390    0.26530  14.338  < 2e-16 ***
factor(age)50              4.60964    0.25984  17.741  < 2e-16 ***
factor(age)55              5.26665    0.25932  20.310  < 2e-16 ***
factor(age)60              5.78588    0.25892  22.346  < 2e-16 ***
```

```
factor(age)65                5.89334    0.25976  22.688  < 2e-16 ***
factor(age)70                6.26493    0.25972  24.122  < 2e-16 ***
factor(age)75                6.59681    0.25978  25.394  < 2e-16 ***
period_1975                  0.29862    0.02770  10.781  < 2e-16 ***
period_1975:age_30_49TRUE   -0.20323    0.07673  -2.649  0.00808 **
period_1975:age_65_79TRUE    0.49290    0.03548  13.891  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 37470.061  on 21  degrees of freedom
Residual deviance:    14.891  on  8  degrees of freedom
AIC: 203.48

Number of Fisher Scoring iterations: 4
```

Answer Using similar notation:

$$
\begin{aligned}
E(D) = \exp\,(&\beta_0 + \beta_1 I(30 \le age < 35) + \beta_2 I(35 \le age < 40) + \\
&\beta_3 I(40 \le age < 45) + \beta_4 I(45 \le age < 50) + \beta_5 I(50 \le age < 55) + \beta_6 I(55 \le age < 60) + \\
&\beta_7 I(60 \le age < 65) + \beta_8 I(65 \le age < 70) + \beta_9 I(70 \le age < 75) + \beta_{10} I(75 \le age < 80) + \\
&\beta_{11} I(1975 \le period < 1980) + \beta_{12} I(1975 \le period < 1980 \& 30 \le age < 50) + \\
&\beta_{13} I(1975 \le period < 1980 \ \&\ 65 \le age < 80) + \log(Y))
\end{aligned}
$$

where β_{11} is the log rate ratio comparing the calendar period 1975-1979 with 1955-1959 for those aged 50-64 years, β_{12} is the log rate ratio comparing the calendar period 1975-1979 with 1955-1959 for those aged less than 50 years with those aged 50-64 years, and β_{12} is the log rate ratio comparing the calendar period 1975-1979 with 1955-1959 for those aged 65 years and over with those aged 50-64 years. *(There is an issue with the interaction terms: those aged less than 30 are lumped in with those aged 50-64 years. I gave additional points for those who found this issue.)*

(e) Using the results from (d), what is the mortality rate ratio and 95% confidence interval for those aged 50-64 years in the 1975-1979 calendar period compared with those aged 50-64 years in the 1955-1959 calendar period? (2 pts)

Answer The mortality rate for those aged 50-64 years in the 1975-1979 calendar period would be $\exp(\beta_0 + [\text{age main effect}] + \beta_{11})$. The mortality rate for those aged 50-64 years in the 1955-1959 calendar period would be $\exp(\beta_0 + [\text{age main effect}])$. The mortality rate ratio of the two groups is then $\exp(\beta_{11}) = \exp(0.29862) =$ `1.348`, with a 95% confidence interval (exp(0.29862-1.96*0.02770),exp(0.29862+1.96*0.02770)) = (`1.277`, `1.423`).

(f) Using the results from (d), show how to calculate the mortality rate ratio for those aged 65-79 years in 1975-1979 calendar period compared with those aged 50-64 years in 1975-1979 calendar period. (2 pts)

Answer *The solution to this question is overly complex and was excluded from the marking.* I then required 6/10 for the other answers in Part 1 (unless someone gave the right answer for this question, for which I gave bonus points). The mortality rate for those aged 65-79 years in the 1975-1979 calendar period would be:

$$\lambda_1 = \exp\left(\beta_0 + \beta_8 I(65 \leq age_1 < 70) + \beta_9 I(70 \leq age_1 < 75) + \right.$$
$$\left. \beta_{10} I(75 \leq age_1 < 80) + \beta_{11} + \beta_{13}\right)$$

where λ_1 is the rate and age_1 is the age. The mortality rate for those aged 50-64 years in the 1975-1979 calendar period would be:

$$\lambda_2 = \exp\left(\beta_0 + \beta_5 I(50 \leq age_2 < 55) + \beta_6 I(55 \leq age_2 < 60) + \right.$$
$$\left. \beta_7 I(60 \leq age_2 < 65) + \beta_{11}\right)$$

where λ_2 is the mortality rate and age_2 is the age. The mortality rate ratio is then

$$\lambda_1/\lambda_2 = \exp\left(\beta_8 I(65 \leq age_1 < 70) + \beta_9 I(70 \leq age_1 < 75) + \right.$$
$$\left. \beta_{10} I(75 \leq age_1 < 80) + \beta_{13}\right) /$$
$$\exp\left(\beta_5 I(50 \leq age_2 < 55) + \beta_6 I(55 \leq age_2 < 60) + \right.$$
$$\left. \beta_7 I(60 \leq age_2 < 65)\right)$$

which is a function of the age in the first age interval and of the age in the second age interval. A simpler revised question would have been: "Using the results from (d), show how to calculate the mortality rate ratio for those aged 65-69 years in 1975-1979 calendar period compared with those aged 50-54 years in 1975-1979 calendar period." The answer to this revised question would have been a mortality rate ratio of $\exp(\beta_{13} + \beta_8 - \beta_5)$.

Part 2

Q2

We now use data from the German Breast Cancer Study Group (GBCSG) on a randomised study of hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients (see `https://doi.org/10.1200/JCO.1994.12.10.2086`). The event considered was time to recurrence of breast cancer or death due to breast cancer ("recurrence-free survival"). The main study found no effect associated with the duration of chemotherapy on recurrence-free survival. The help page for the dataset is shown below:

```
library(rstpm2) # brcancer, stpm2
help("brcancer", help_type="text")

German breast cancer data from Stata.

Description:

     See <URL: https://www.stata-press.com/data/r11/brcancer.dta>.

Usage:

     data(brcancer)

Format:

     A data frame with 686 observations on the following 15 variables.

     'id' a numeric vector
```

‘hormon’ hormonal therapy

‘x1’ age, years

‘x2’ menopausal status

‘x3’ tumour size, mm

‘x4’ tumour grade

‘x5’ number of positive nodes

‘x6’ progesterone receptor, fmol

‘x7’ estrogen receptor, fmol

‘rectime’ recurrence free survival time, days

‘censrec’ censoring indicator

‘x4a’ tumour grade>=2

‘x4b’ tumour grade==3

‘x5e’ exp(-0.12*x5)

We now define the event time as the time from randomisation to time of recurrence or death – that is, we are modelling for recurrence-free survival. There were 299 events and the event times are in days from randomisation.

(a) The Kaplan-Meier estimators for the survival functions by tumour grade are shown in Figure 2. Carefully describe and interpret the three survival curves. (2 pts)

```
library(survival) # survfit, survdiff, Surv, coxph, cox.zph
sfit = survfit(Surv(rectime, censrec==1)~x4, data=brcancer)
plot(sfit, col=1:3, xlab="Recurrence free survival time (days)", ylab="Survival")
legend("topright", paste("x4 =", 1:3), col=1:3, lty=1, bty="n")
```

Answer The black curve is the estimated survival for those tumour grade 1. There are comparatively few in this group, as indicated by the large steps in survival. Survival is approximately 0.8 at 1000 days. The red curve is survival for tumour grade 2. There are more at risk, as indicated by the small steps in the survival curve. Survival is approximately 0.65 at 1000 days, with a median survival of approximately 1500 days. Tumour grade 2 has poorer survival than tumour grade 1. There are fewer at risk after 2000 days. The green curve is survival for tumour grade 3. There are fewer at risk than tumour grade 2. Survival is generally poorer for tumour grade 3 compared with tumour grade 2, although the curves intersect at close to 2500 days. Survival for tumour grade 3 is approximately 0.58 at 1000 days, with a median survival of approximately 1200 days.

(b) We now use a log-rank test to compare the three curves. What is the null hypothesis for the log-rank test? Based on the output, what can we conclude? (2 pts)

```
survdiff(Surv(rectime, censrec==1)~x4, data=brcancer)
```

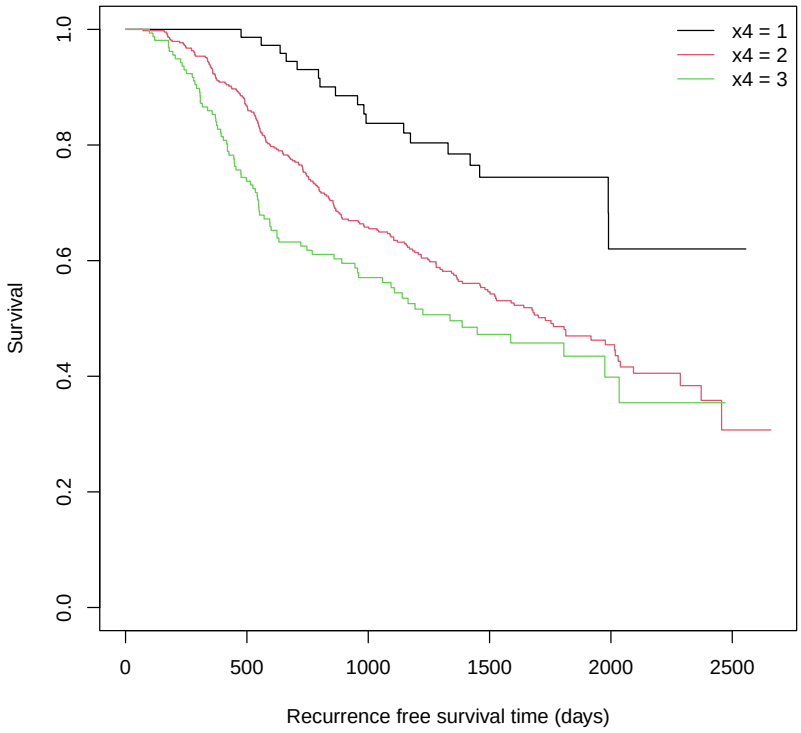

Figure 2: Kaplan-Meier survival curves by tumour grade, German Breast Cancer Study Group

```
Call:
survdiff(formula = Surv(rectime, censrec == 1) ~ x4, data = brcancer)

       N Observed Expected (O-E)^2/E (O-E)^2/V
x4=1  81       18     42.2   13.8469    16.159
x4=2 444      202    198.2    0.0725     0.215
x4=3 161       79     58.6    7.0788     8.848

 Chisq= 21.1  on 2 degrees of freedom, p= 3e-05
```

Answer The null hypothesis is that there are no differences between the three survival curves. From the p-value of 0.00003, we can reject the null hypothesis - we have reasonably strong evidence for survival differences between the three groups. Given the observed and expected values, we could interpret that survival would tend to be better for tumour grade 1 cancers, and poorer for tumour grade 3 cancers.

(c) Write out the regression equation for the first Cox model specified in the following code. (2 pts)

```
brcancer2 = transform(brcancer, x4_2 = ifelse(x4==2,1,0),
      x4_3 = ifelse(x4==3,1,0))
fit = coxph(Surv(rectime,censrec==1)~hormon+x4_2+x4_3, data=brcancer2)
summary(fit)
cat("\n") # add a newline to separate the summary and the anova
fit0 = coxph(Surv(rectime,censrec==1)~hormon, data=brcancer2)
anova(fit0, fit)
```

```
Call:
coxph(formula = Surv(rectime, censrec == 1) ~ hormon + x4_2 +
    x4_3, data = brcancer2)

  n= 686, number of events= 299

          coef exp(coef) se(coef)      z Pr(>|z|)
hormon -0.3404    0.7115   0.1254 -2.714 0.006651 **
x4_2    0.8724    2.3927   0.2461  3.546 0.000392 ***
x4_3    1.1247    3.0792   0.2617  4.297 1.73e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

       exp(coef) exp(-coef) lower .95 upper .95
hormon    0.7115     1.4056    0.5564    0.9098
x4_2      2.3927     0.4179    1.4773    3.8756
x4_3      3.0792     0.3248    1.8435    5.1432

Concordance= 0.594  (se = 0.016 )
Likelihood ratio test= 31.88  on 3 df,   p=6e-07
Wald test            = 27.04  on 3 df,   p=6e-06
Score (logrank) test = 28.48  on 3 df,   p=3e-06

Analysis of Deviance Table
 Cox model: response is  Surv(rectime, censrec == 1)
 Model 1: ~ hormon
 Model 2: ~ hormon + x4_2 + x4_3
   loglik  Chisq Df P(>|Chi|)
1 -1783.7
2 -1772.2 23.057  2 9.845e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Answer using our previous notation, we have that

$$h(t|x) = h_0(t) \exp\left(\beta_1 hormon + \beta_2 I(x4 = 2) + \beta_3 I(x4 = 3)\right)$$

where $h(t|x)$ is the hazard at time t given covariates x (in this case, hormone treatment and tumour grade), $h_0(t)$ is the baseline hazard function, β_1 is the log hazard ratio for hormone treatment, β_2 is the log hazard ratio comparing tumour grade 2 with tumour grade 1, and β_3 is the log hazard ratio comparing tumour grade 3 with tumour grade 1.

(d) Based on the previous output, discuss whether there is any evidence that tumour grade is associated with recurrence-free survival. Provide confidence intervals and p-values to support your argument. (2 pts)

Answer After adjustment for hormone treatment and time since study entry, we find that the p-values for tumour grades 2 and 3 compared with tumour grade 1 are 0.0004 and 0.00002, respectively. Both of these p-values indicate strong associations. Moreover, the likelihood ratio test in the `anova` output gives a p-value of 0.00001 for the two tumour grade parameters - again supporting a strong association. The estimated hazards ratio and 95% confidence interval (CI) for tumour grade 2 compared with tumour grade 1 is

2.393 (95% CI: 1.477, 3.876). The estimated hazards ratio and 95% confidence interval (CI) for tumour grade 3 compared with tumour grade 1 is 3.079 (95% CI: 1.844, 5.143).

(e) Based on the following Schoenfeld residuals tables, is there any evidence for non-proportionality in the modelled covariates? Interpret the tables and explain your reasoning. (2 pts)

```
cox.zph(fit)
## do again using x4 as a factor (rather than as indicator variables)
cox.zph(coxph(Surv(rectime,censrec==1)~hormon+factor(x4), data=brcancer))
```

```
        chisq df      p
hormon  0.194  1 0.6598
x4_2    3.649  1 0.0561
x4_3   10.433  1 0.0012
GLOBAL 13.208  3 0.0042
            chisq df      p
hormon      0.194  1 0.6598
factor(x4) 13.153  2 0.0014
GLOBAL     13.208  3 0.0042
```

Answer The null hypothesis for the global test is that none of the covariates have non-proportional hazard ratios (assuming a log-linear alternative hypothesis). Given the p-value of 0.004, we find evidence for non-proportionality. From the second table, we find evidence that tumour grade is non-proportional, while the first table suggests borderline evidence for `x4_2` and strong evidence for `x4_3`.

(f) Based on the following plot of Schoenfeld residuals (Figure 3), how would you expect the hazard ratio for tumour grade 3 compared with tumour grade 1 to vary by time since randomisation? Explain your reasoning. (2 pts)

```
plot(cox.zph(fit)[3])
```

Answer The smoothed Schoenfeld residuals suggests a log hazard ratio of approximately 2.5 at 270 days, with an attenuated log hazard ratio to below 0 at around 1000 days, with either a constant or rising log hazard ratio from 1400 days. There are fewer events at the end of follow-up, so that the confidence intervals tend to form a "trumpet".

(g) We now fit a Cox regression model adjusting for **hormon**, $\mathbf{x4_2}$ and $\mathbf{x4_3}$, with a time-varying effect for $\mathbf{x4_3}$ when **rectime**$\geq$ 1000 (see the following output). Give a formula for this regression model. Is there any evidence for a time-varying effect? What is the form of the time-varying hazard ratio? (3 pts)

```
brcancer3 <- survSplit(brcancer2, cut=c(0,1000,10000), end="rectime", start="start",
  event="censrec")
fit2 <- coxph(Surv(start,rectime,censrec==1)~hormon+x4_2+x4_3+I(x4_3 & start>=1000),
      data=brcancer3)
summary(fit2)
```

```
Call:
coxph(formula = Surv(start, rectime, censrec == 1) ~ hormon +
    x4_2 + x4_3 + I(x4_3 & start >= 1000), data = brcancer3)

  n= 1038, number of events= 299
```

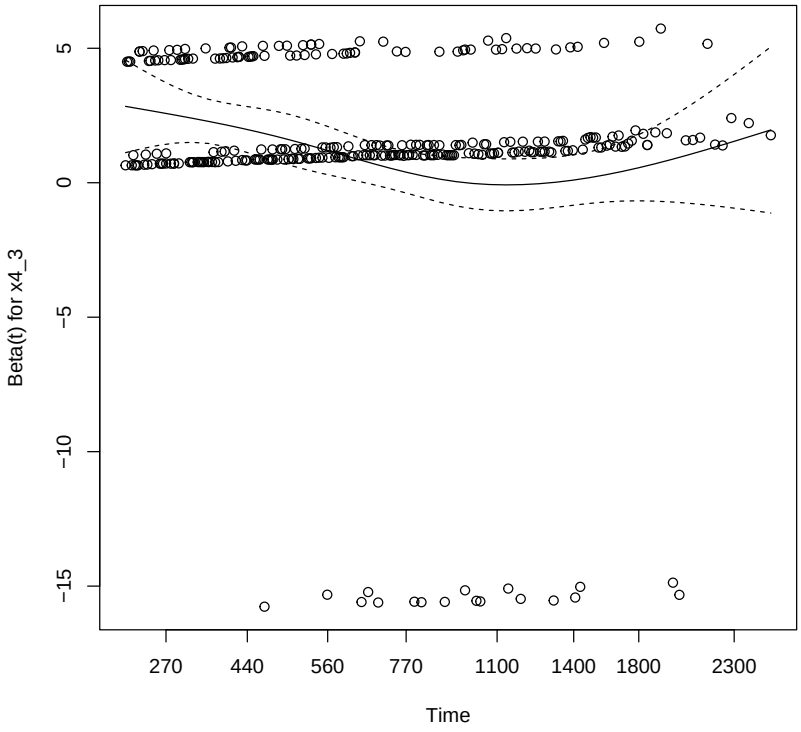

Figure 3: Schoenfeld residual plot for tumour grade 3 compared with tumour grade 1, German Breast Cancer Study Group

```
                                  coef exp(coef) se(coef)      z Pr(>|z|)
hormon                         -0.3423    0.7101   0.1255 -2.728 0.006364 **
x4_2                            0.8760    2.4014   0.2461  3.560 0.000371 ***
x4_3                            1.2403    3.4566   0.2713  4.572 4.83e-06 ***
I(x4_3 & start >= 1000)TRUE -0.4951    0.6095   0.3294 -1.503 0.132836
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                            exp(coef) exp(-coef) lower .95 upper .95
hormon                         0.7101     1.4083    0.5553    0.9081
x4_2                           2.4014     0.4164    1.4825    3.8896
x4_3                           3.4566     0.2893    2.0311    5.8826
I(x4_3 & start >= 1000)TRUE    0.6095     1.6406    0.3196    1.1624

Concordance= 0.604  (se = 0.016 )
Likelihood ratio test= 34.28  on 4 df,   p=7e-07
Wald test            = 29.64  on 4 df,   p=6e-06
Score (logrank) test = 31.22  on 4 df,   p=3e-06
```

Answer Using our previous notation, we have that

$$h(t|x) = h_0(t) \exp\left(\beta_1 hormon + \beta_2 I(x4 = 2) + \beta_3 I(x4 = 3) + \beta_4 I(x4 = 3 \ \& \ t \geq 1000)\right)$$

where β_4 is the log hazard ratio for tumour grade 3 compared with tumour grade 1 for the time period after 1000 days compared with the time period less than 1000 days. The p-value for this time-varying effect (or interaction) is not significant (p=0.13). The estimated hazard ratio is 0.610 (95% CI: 0.320, 1.162), which suggests a reduction that is not statistically significant. A function for the hazard ratio comparing tumour grade 3 with tumour grade 1 is $HR(t) = \exp(\beta_3 + \beta_4 I(t \geq 1000))$.

(h) We now fit a time-varying effect which is linear in time. Give the estimated equation for the hazard ratio for tumour grade = 3 compared with tumour grade = 1 by time. (2 pts)

```
fit3 = coxph(Surv(start,rectime,censrec==1)~hormon+x4_2+x4_3+tt(x4_3), data=brcancer3,
     tt=function(x,t,...) x*t)
summary(fit3)

Call:
coxph(formula = Surv(start, rectime, censrec == 1) ~ hormon +
    x4_2 + x4_3 + tt(x4_3), data = brcancer3, tt = function(x,
    t, ...) x * t)

  n= 1038, number of events= 299

               coef  exp(coef)   se(coef)      z Pr(>|z|)
hormon   -0.3439430  0.7089693  0.1254947 -2.741 0.006131 **
x4_2      0.8811904  2.4137714  0.2460729  3.581 0.000342 ***
x4_3      1.7870555  5.9718424  0.3468534  5.152 2.57e-07 ***
tt(x4_3) -0.0009257  0.9990747  0.0003376 -2.742 0.006112 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
          exp(coef) exp(-coef) lower .95 upper .95
hormon       0.7090     1.4105    0.5544    0.9067
x4_2         2.4138     0.4143    1.4902    3.9098
x4_3         5.9718     0.1675    3.0260   11.7856
tt(x4_3)     0.9991     1.0009    0.9984    0.9997

Concordance= 0.609  (se = 0.015 )
Likelihood ratio test= 40.51  on 4 df,   p=3e-08
Wald test            = 35.93  on 4 df,   p=3e-07
Score (logrank) test = 37.88  on 4 df,   p=1e-07
```

Answer Although this was not asked for, I will first give the regression equation:

$$h(t|x) = h_0(t) \exp\left(\beta_1 hormon + \beta_2 I(x4 = 2) + \beta_3 I(x4 = 3) + \beta_4 t I(x4 = 3)\right)$$

From this we see that the hazard ratio for tumour grade 3 compared with tumour grade 1 is $\exp(\beta_3 + \beta_4 t) = \exp(1.787 - 0.000926t)$. When $t = 0$, the hazard ratio is 5.972 (95% CI 3.026, 11.786), which is highly significant. When $t = 1000$, the hazard ratio is `2.366`. The p-value for the time-varying effect (0.006) is statistically significant.

Q3

(a) Compare and contrast (i) Poisson regression models, (ii) Cox regression models and (iii) flexible parametric survival models. Summarise the advantages and disadvantages of each model of the three models. For each of the three models, describe an example study that you would use for that model and explain why you would use that model over the other two models. (3 pts)

Answer

All three models are proportional hazards models (although the flexible parametric model can also be used for additive hazards, proportional odds or probit models). All models allow for left truncated and right censored data. All models allow for time-varying effects, and they can all model for piecewise-constant time-varying covariates.

Model	Advantages	Disadvantages
Cox regression	- Non-parametric baseline	- Difficult to model time-varying effects using splines - Unable to model for interval-censored data
Poisson regression	- Easily accommodates multiple time scales	- Requires time-splitting to adjust for time - Unable to model for interval-censored data - Parametric baseline
Flexible parametric	- Easily accommodates splines for time-varying effects - Allows for interval-censored data	- Parametric baseline - Issue with interpretting multiple time-varying effects

The estimating a time-constant hazard ratio with a single time scale, I would use Cox regression, as it allows for a non-parametric baseline. An example of a typical study would be a randomised controlled trial for baseline treatment. For estimating a single smooth time-varying effect with a single time scale, I would use flexible parametric survival models. I would also use this model class for estimating survival differences and for standardised estimators. An example study would be a register-based study. For estimating more than one smooth time-varying effect or more than one time scale, I would use Poisson regression. I would also use Poisson regression with aggregated event and person-time data. An example study would be a register-based study.

(b) What is the relationship between analysing a cohort study using Cox regression and analysing a nested case-control study using conditional logistic regression? Describe this relationship in terms of risk sets. (1 pt)

Answer A Cox regression is analysed using a partial likelihood. A nested case-control study is analysed using a conditional logistic regression, which is mathematically equivalent to the partial likelihood except that the risk set in the denominator is a *sample* of the full risk set.

(c) Design a cohort study to investigate the effect of smoking cessation on lung cancer mortality. Describe the design in terms of: eligibility criteria; time scales; how events are defined; how end of follow-up is defined; primary outcome measures; and analysis methods. (4 pts)

Answer This could be answered in terms of a healthy population or a cohort of lung cancer patients.

First, I consider a health population. Eligibility: current smokers who have not been diagnosed with lung cancer. We could exclude those with related diseases (e.g. chronic obstructive pulmonary disease) or respiratory symptoms. Time scales: time since cessation (for those who cease) and attained age are obvious time scales. For smoking, a common exposure measure would be pack-years of smoking. The primary event will death due to lung cancer. Follow-up will be until the earliest of a fixed date for the end of study, censoring (e.g. migration) or the date of death. For analysis, we would split person-time by attained age and time since cessation. This could be done using Poisson regression with splines for the two time scales. We would also need to adjust for pack-years of smoking exposure. As a technical issue, observed smoking exposure may be due to symptoms, where lung cancer incidence (and mortality) may actually rise soon after cessation. A further technical challenge is that smoking "cessation" may not be ongoing, where a former smoker may re-start smoking. This leads to further analytical challenges!

Second, I consider a cohort of lung cancer patients. Eligibility: diagnosed with lung cancer and ever having smoked. Time scales: time from the cancer diagnosis and time from smoking cessation are obvious time scales. Primary event: death due to cancer. Follow-up will the until the earliest of a fixed date for the end of study, fix follow-up period (e.g. five years), censoring (e.g. migration) or date of death. As per the population case, we would use a time-dependent exposure for smoking, splitting person-time by two time scales. Again, I would use Poisson regression to model for multiple time scales using splines.

(Part 1: 12 pts; Part 2: 25 pts)