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

Mark Clements

8–17 November, 2021

Contents

Instructions

- **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")
```

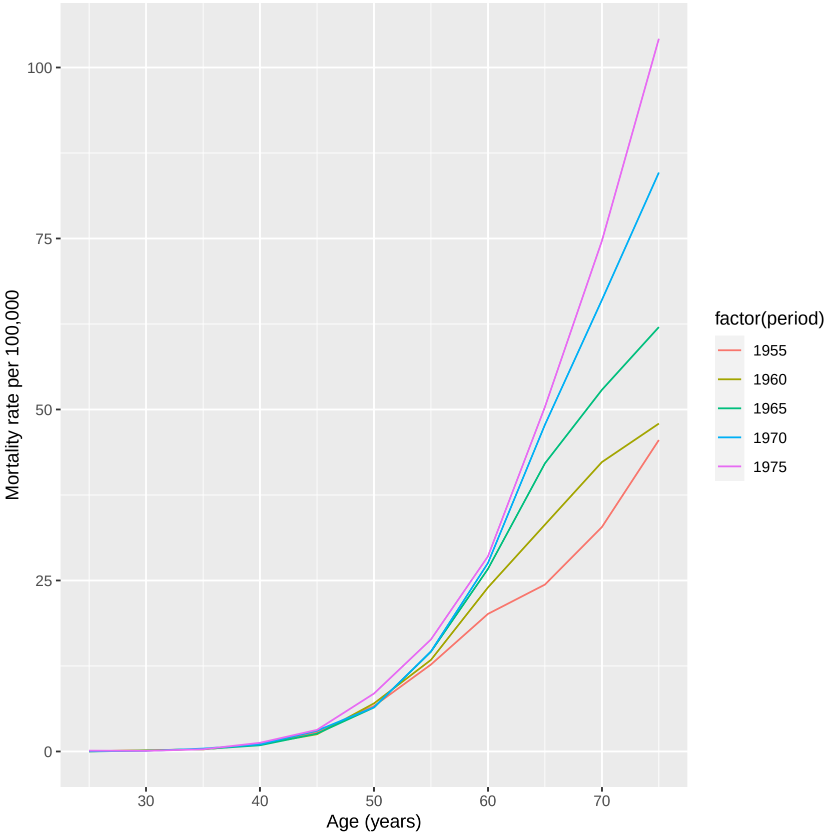

Figure 1: Age-specific bladder cancer mortality rates by calendar period, males, Italy, 1955-1979

(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
```

(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)

(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
```

(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)

(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)

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")
```

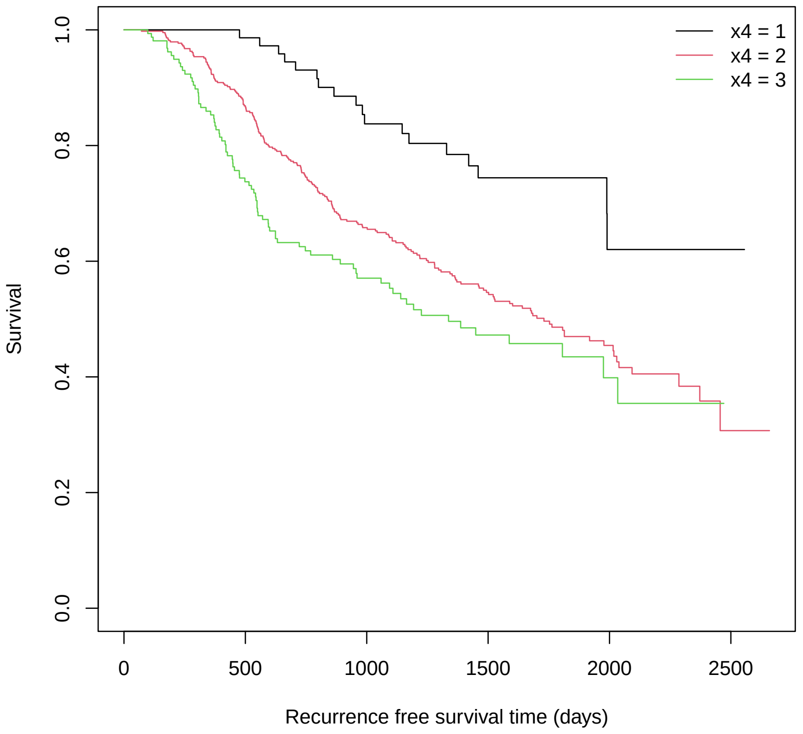

Figure 2: Kaplan-Meier survival curves by tumour grade, German Breast Cancer Study Group

(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)

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
```

(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
```

(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)

(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
```

(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])
```

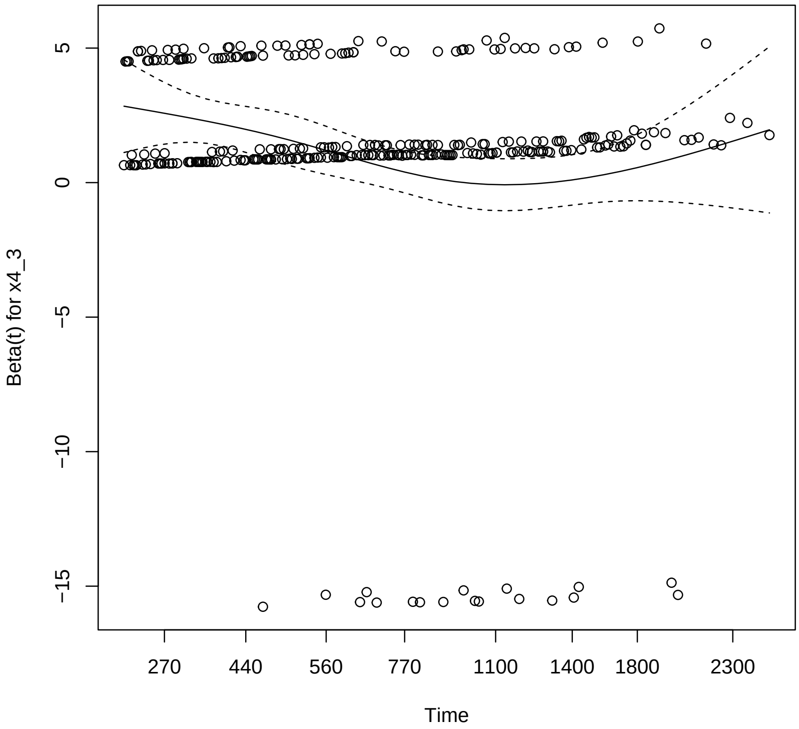

Figure 3: Schoenfeld residual plot for tumour grade 3 compared with tumour grade 1, German Breast Cancer Study Group

(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

                               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
```

(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
```

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)

(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)

(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)

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