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

Mark Clements

7–16 November, 2022

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 16 November 2022 and **the examination is due by 17:00 on Wednesday 23 November 2022**.
- The examination will be graded and results returned to you by Wednesday 1 December 2022.
- The examination is in two parts. To pass the examination, you need to score at least 8/13 for Part 1 focused on rates and general regression modelling and 12/21 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

Thun et al (1997; `https://cancercontrol.cancer.gov/sites/default/files/2020-08/m08_4.pdf`) report on results from two large US population-based cohorts: CPS-I from 1959–1965 and CPS-II from 1982–1988. The authors report the number of lung cancers (`y`) and person-year exposed (`pt`) broken down by:

- Cohort study, represented using the variable `cpsii`, which is 1 for CPS-II and 0 for CPS-I
- Sex, which is encoded using the variable `male`, which is 1 for males and 0 for females
- Attained age, which is encoded in the variable `age` for the lower bound of the five-year age groups 50–54, 55–59, . . . , 75–79 years

- Smoking status, this is encoded in the variable `smoker`, which is 1 for current smokers and 0 for never smokers (that is, former smokers are not presented).

These data are represented as data-frame named `df`.

```
str(df)

’data.frame’: 48 obs. of  7 variables:
 $ y     : int  6 13 16 14 13 8 20 23 38 34 ...
 $ pt    : num  105263 95588 76190 60606 43771 ...
 $ age   : int  50 55 60 65 70 75 50 55 60 65 ...
 $ male  : int  1 1 1 1 1 1 0 0 0 0 ...
 $ cpsii : int  0 0 0 0 0 0 0 0 0 0 ...
 $ smoker: int  0 0 0 0 0 0 0 0 0 0 ...
 $ female: int  0 0 0 0 0 0 1 1 1 1 ...
```

Q1

(a) We fit a Poisson regression model for lung cancer mortality rates in CPS-II, with a main effect for attained age as a factor and a two-way interaction between being male, and being a current smoker (compared with being a never smoker). Write a formula for this regression model. As a reminder, please define your notation. (2 pts)

```
fit <- glm(y~factor(age)+male*smoker+offset(log(pt)),
   data=df, subset=(cpsii==1),family=poisson)
summary(fit)


Call:
glm(formula = y ~ factor(age) + male * smoker + offset(log(pt)),
    family = poisson, data = df, subset = (cpsii == 1))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.79012  -0.81503  -0.02832   0.77967   1.94115

Coefficients:
               Estimate Std. Error  z value Pr(>|z|)
(Intercept)   -10.07191    0.08988 -112.064  < 2e-16 ***
factor(age)55   0.55717    0.07797    7.146 8.91e-13 ***
factor(age)60   1.05863    0.07397   14.311  < 2e-16 ***
factor(age)65   1.52816    0.07327   20.858  < 2e-16 ***
factor(age)70   1.88285    0.07547   24.947  < 2e-16 ***
factor(age)75   2.09603    0.08483   24.709  < 2e-16 ***
male            0.13933    0.12085    1.153    0.249
smoker          2.63805    0.07283   36.223  < 2e-16 ***
male:smoker     0.62558    0.12754    4.905 9.34e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 5817.22  on 23  degrees of freedom
```

```
Residual deviance:   29.82  on 15  degrees of freedom
AIC: 192.76

Number of Fisher Scoring iterations: 4
```

(b) Our research question relates to how the rate ratios for current versus never smokers varies by cohort study and by sex. From the regression model, show how to calculate (either algebraically or numerically) the lung cancer mortality rate ratio for (i) females in CPS-II and (ii) males in CPS-II. (4 pts)

(c) From the output above, calculate the smoking rate ratio and 95% confidence interval for females in CPS-II. (2 pts)

(d) Based on this model fit, show R computer code to calculate the smoking rate ratio and 95% confidence intervals for males in CPS-II . (2 pts)

(e) For the following output, we used the same model but for the CPS-I cohort study. Describe a method to assess whether the smoking rate ratio for females in CPS-II is significantly higher than that for females in CPS-I. *(Hint: not necessarily based on the available output.)* (3 pts)

```
fit <- glm(y~factor(age)+male*smoker+offset(log(pt)),
   data=df, subset=(cpsii==0),family=poisson)
summary(fit)


Call:
glm(formula = y ~ factor(age) + male * smoker + offset(log(pt)),
    family = poisson, data = df, subset = (cpsii == 0))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.57337  -0.57215  -0.03735   0.49529   2.16772

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -10.16013    0.10287 -98.771  < 2e-16 ***
factor(age)55   0.44491    0.08957   4.967 6.79e-07 ***
factor(age)60   1.05330    0.08719  12.081  < 2e-16 ***
factor(age)65   1.37029    0.09216  14.869  < 2e-16 ***
factor(age)70   1.52781    0.10707  14.269  < 2e-16 ***
factor(age)75   1.87375    0.12759  14.685  < 2e-16 ***
male            0.51504    0.14226   3.620 0.000294 ***
smoker          1.25371    0.12173  10.299  < 2e-16 ***
male:smoker     1.29623    0.17263   7.509 5.98e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2297.187  on 23  degrees of freedom
Residual deviance:   18.125  on 15  degrees of freedom
AIC: 160.15
```

```
Number of Fisher Scoring iterations: 4
```

Part 2

Q2

(a) Given the following data, calculate the Kaplan-Meier and actuarial estimators for all cause survival at time 2 years. For the actuarial estimator, assume one-year time bands. Please show your working. *(Hint: you can either do this by hand or use software.)* (2 pts)

Last time observed (months)	Status at last observation
2	Alive
4	Death
5	Alive
9	Death
11	Death
15	Alive
15	Death
15	Alive
19	Alive
23	Death
23	Death
25	Alive
26	Death

Q3

We now use data on chemotherapy for stage B/C colon cancer represented in the `survival::colon` dataset. The help page for the dataset is shown below:

```
library(survival)
help("colon", help_type="text", package="survival")
```

```
colon                  package:survival                  R Documentation

Chemotherapy for Stage B/C colon cancer

Description:

     These are data from one of the first successful trials of adjuvant
     chemotherapy for colon cancer. Levamisole is a low-toxicity
     compound previously used to treat worm infestations in animals;
     5-FU is a moderately toxic (as these things go) chemotherapy
     agent. There are two records per person, one for recurrence and
     one for death

Usage:

     colon
            data(cancer, package="survival")

Format:
       id:        id
       study:     1 for all patients
       rx:        Treatment - Obs(ervation), Lev(amisole), Lev(amisole)+5-FU
       sex:       1=male
       age:       in years
       obstruct:  obstruction of colon by tumour
       perfor:    perforation of colon
       adhere:    adherence to nearby organs
       nodes:     number of lymph nodes with detectable cancer
       time:      days until event or censoring
       status:    censoring status
       differ:    differentiation of tumour (1=well, 2=moderate, 3=poor)
       extent:    Extent of local spread (1=submucosa, 2=muscle, 3=serosa,
                  4=contiguous structures)
       surg:      time from surgery to registration (0=short, 1=long)
       node4:     more than 4 positive lymph nodes
       etype:     event type: 1=recurrence,2=death

Note:

     The study is originally described in Laurie (1989).  The main
     report is found in Moertel (1990).  This data set is closest to
     that of the final report in Moertel (1991).

References:

     JA Laurie, CG Moertel, TR Fleming, HS Wieand, JE Leigh, J Rubin,
     GW McCormack, JB Gerstner, JE Krook and J Malliard.  Surgical
     adjuvant therapy of large-bowel carcinoma: An evaluation of
     levamisole and the combination of levamisole and fluorouracil: The
     North Central Cancer Treatment Group and the Mayo Clinic.  J
     Clinical Oncology, 7:1447-1456, 1989.

     CG Moertel, TR Fleming, JS MacDonald, DG Haller, JA Laurie, PJ
     Goodman, JS Ungerleider, WA Emerson, DC Tormey, JH Glick, MH
     Veeder and JA Maillard.  Levamisole and fluorouracil for adjuvant
     therapy of resected colon carcinoma. New England J of Medicine,
     332:352-358, 1990.

     CG Moertel, TR Fleming, JS MacDonald, DG Haller, JA Laurie, CM
     Tangen, JS Ungerleider, WA Emerson, DC Tormey, JH Glick, MH Veeder
     and JA Maillard, Fluorouracil plus Levamisole as an effective
     adjuvant therapy after resection of stage II colon carcinoma: a
     final report.  Annals of Internal Med, 122:321-326, 1991.
```

(a) The Kaplan-Meier estimators for the survival functions by treatment arm are shown in Figure 1. Carefully describe and interpret the three survival curves. (2 pts)

```
sfit = survfit(Surv(time, status)~rx, data=survival::colon, subset=(etype==1))
plot(sfit, col=1:3, xlab="Recurrence-free survival time (days)", ylab="Survival")
legend("topright", paste("rx =", levels(survival::colon$rx)), col=1:3, lty=1, bty="n")
```

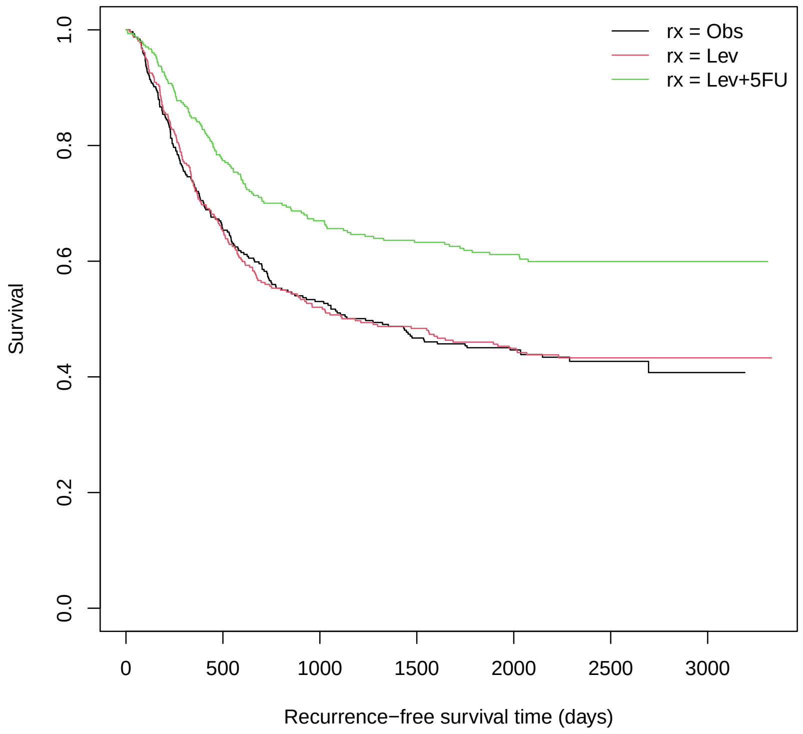

Figure 1: Kaplan-Meier survival curves by treatment arm

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

```
colon2 = transform(subset(survival::colon, etype==1),
    Lev = ifelse(rx=="Lev",1,0),
    Lev_5FU = ifelse(rx=="Lev+5FU",1,0))
fit = coxph(Surv(time,status)~Lev+Lev_5FU+factor(extent), data=colon2)
summary(fit)

Call:
coxph(formula = Surv(time, status) ~ Lev + Lev_5FU + factor(extent),
    data = colon2)

  n= 929, number of events= 468

                    coef exp(coef) se(coef)      z Pr(>|z|)
Lev             -0.02149   0.97874  0.10738 -0.200  0.84138
Lev_5FU         -0.51040   0.60025  0.11883 -4.295 1.75e-05 ***
factor(extent)2  0.26937   1.30915  0.47957  0.562  0.57432
factor(extent)3  0.95936   2.61003  0.45069  2.129  0.03328 *
factor(extent)4  1.49213   4.44654  0.48494  3.077  0.00209 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                exp(coef) exp(-coef) lower .95 upper .95
Lev                0.9787     1.0217    0.7930    1.2080
```

```
Lev_5FU            0.6003     1.6660    0.4755    0.7577
factor(extent)2    1.3091     0.7639    0.5114    3.3512
factor(extent)3    2.6100     0.3831    1.0790    6.3135
factor(extent)4    4.4465     0.2249    1.7189   11.5028

Concordance= 0.589  (se = 0.013 )
Likelihood ratio test= 56.48  on 5 df,   p=6e-11
Wald test            = 51.33  on 5 df,   p=7e-10
Score (logrank) test = 53.5  on 5 df,   p=3e-10
```

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

(d) 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)

               chisq df    p
Lev            0.075  1 0.78
Lev_5FU        0.232  1 0.63
factor(extent) 1.161  3 0.76
GLOBAL         1.394  5 0.92
```

(e) Based on the following plot of Schoenfeld residuals (Figure 2), how would you expect the hazard ratio for the Lev+5FU arm compared with the Observation arm to vary by time since randomisation? Explain your reasoning. (2 pts)

```
plot(cox.zph(fit)[2])
```

Q4

For the following research questions:

- Specify a target estimand (that is, the parameter that you want to estimate to address the research question)
- Describe a prospective study design, including target population, the event outcome of interest and any censoring
- Select and justify the choice of regression model to estimate the target estimand
- Carefully describe the time scale or time scales used for modelling and motivate your choice of time scale(s)
- Consider whether and how to adjust for potential confounding
- Consider whether and how to adjust for potential time-varying effects

The research questions are:

(a) How does lung cancer mortality differ between never and former smokers? How does time since smoking cessation affect mortality? (3 pts)

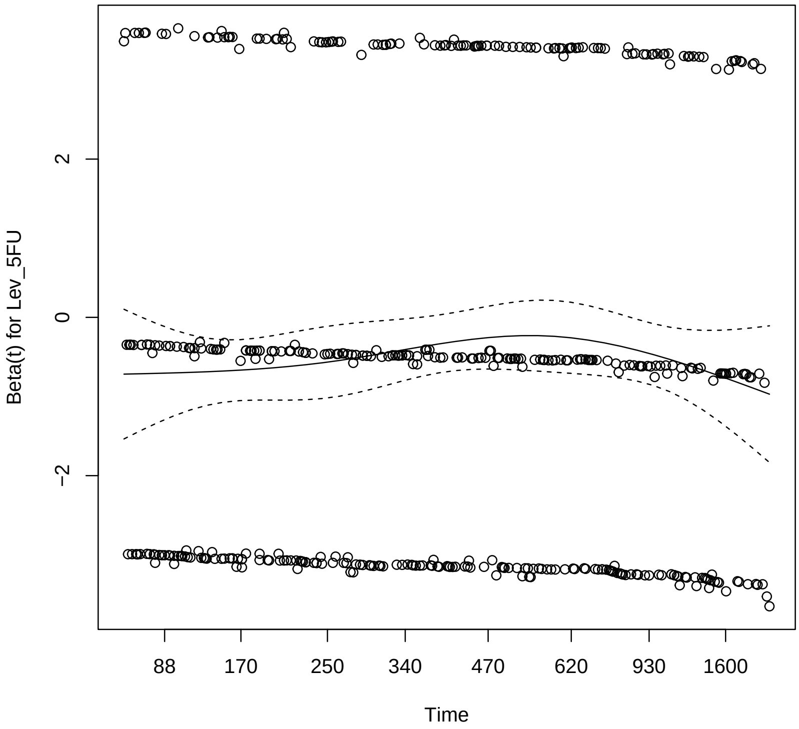

Figure 2: Schoenfeld residual plot for comparing the Lev+5FU arm with the Observation arm

(b) Does human papillomavirus (HPV) vaccination reduce cervical cancer incidence? (3 pts)

(c) For men diagnosed with prostate cancer, does initial treatment assignment to radical prostatectomy (removal of the prostate) improve all cause survival compared with initial treatment assignment to radiation therapy? (3 pts)

(Part 1: 13 pts; Part 2: 21 pts)