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

Mark Clements

4–13 November, 2024

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 iThenticate in order to assess potential plagiarism.
- The examination will be made available by noon on Wednesday 13 November 2024 and **the examination is due by 17:00 on Wednesday 20 November 2024**.
- The examination will be graded and results returned to you by Wednesday 27 November 2024.
- The examination is in two parts. To pass the examination, you need to score at least 9/17 for Part 1 focused on rates and general regression modelling and 13/24 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

The **survival** package on CRAN includes the `colon` dataset which is follow-up from a randomised controlled trial of three different treatment modalities for male colon cancer patients. We include a subset of the dataset (named `colon_recurrance`) restricted to the recurrence times with the following variables:

rx Treatment – Obs(ervation), Lev(amisole), Lev(amisole)+5-FU

age in years

differ differentiation of tumour (1=well, 2=moderate, 3=poor)

time days until event or censoring

status censoring status

```
summary(colon_recurrence)
```

```
      rx           age            differ           time          status
Obs    :315   Min.   :18.00   Min.   :1.000   Min.   :   8   Min.   :0.0000
Lev    :310   1st Qu.:53.00   1st Qu.:2.000   1st Qu.: 370   1st Qu.:0.0000
Lev+5FU:304   Median :61.00   Median :2.000   Median :1548   Median :1.0000
              Mean   :59.75   Mean   :2.063   Mean   :1405   Mean   :0.5038
              3rd Qu.:69.00   3rd Qu.:2.000   3rd Qu.:2289   3rd Qu.:1.0000
              Max.   :85.00   Max.   :3.000   Max.   :3329   Max.   :1.0000
                              NA's   :23
```

Q1

(a) We fit a Poisson regression model for the time from study entry to recurrence or death adjusting for age, treatment and differentiation (see code below). Write a formula for this regression model. As a reminder, please define your notation. (4 pts)

```
fit <- glm(status~I(age-60)+I(rx=="Lev")+I(rx=="Lev+5FU")+I(differ==2)+I(differ==3)+
               offset(log(time/365.25)),
           data=colon_recurrence, family=poisson)
summary(fit)
```

```
Call:
glm(formula = status ~ I(age - 60) + I(rx == "Lev") + I(rx ==
    "Lev+5FU") + I(differ == 2) + I(differ == 3) + offset(log(time/365.25)),
    family = poisson, data = colon_recurrence)

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            -1.969016   0.165588 -11.891  < 2e-16 ***
I(age - 60)            -0.006299   0.003985  -1.580  0.11399
I(rx == "Lev")TRUE     -0.043578   0.108705  -0.401  0.68850
I(rx == "Lev+5FU")TRUE -0.593727   0.119418  -4.972 6.63e-07 ***
I(differ == 2)TRUE      0.068604   0.160959   0.426  0.66995
I(differ == 3)TRUE      0.559015   0.185161   3.019  0.00254 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1839.0  on 905  degrees of freedom
Residual deviance: 1790.5  on 900  degrees of freedom
  (23 observations deleted due to missingness)
AIC: 2718.5

Number of Fisher Scoring iterations: 7
```

(b) From the regression model output, carefully explain how to interpret the six estimated parameters. (3 pts)

(c) From the regression model output, show how to numerically calculate the rate ratio for treatment for Lev+5FU compared with observed treatment, including the 95% confidence interval. (2 pts)

(d) From the regression model output, show how to numerically calculate the predicted rate for patients aged 60 years in the observed treatment group with a well differentiated tumour, including the 95% confidence interval. (2 pts)

(e) From the regression model output, show how to numerically calculate the predicted rate for patients aged 70 years with poorly differentiated tumours in the Lev+5FU treatment arm. (2 pts)

(f) We now extend the regression model output to include an interaction between `I(rx=="Lev+5FU")` and `I(differ==2)` and between `I(rx=="Lev+5FU")` and `I(differ==3)` (see R code and output below). Explain how to interpret the main effect for `I(rx=="Lev+5FU")` for this model. (2 pts)

```
fit2 <- glm(status~I(age-60)+I(rx=="Lev")+I(rx=="Lev+5FU")+I(differ==2)+I(differ==3)+
               I(rx=="Lev+5FU"):I(differ==2)+
               I(rx=="Lev+5FU"):I(differ==3)+
               offset(log(time/365.25)),
          data=colon_recurrence, family=poisson)
summary(fit2)
```

```
Call:
glm(formula = status ~ I(age - 60) + I(rx == "Lev") + I(rx ==
    "Lev+5FU") + I(differ == 2) + I(differ == 3) + I(rx == "Lev+5FU"):I(differ ==
    2) + I(rx == "Lev+5FU"):I(differ == 3) + offset(log(time/365.25)),
    family = poisson, data = colon_recurrence)

Coefficients:
                                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                                 -1.891940   0.181263 -10.438   <2e-16 ***
I(age - 60)                                 -0.006225   0.003991  -1.560   0.1188
I(rx == "Lev")TRUE                          -0.047008   0.108798  -0.432   0.6657
I(rx == "Lev+5FU")TRUE                      -0.917953   0.379323  -2.420   0.0155 *
I(differ == 2)TRUE                          -0.006427   0.181204  -0.035   0.9717
I(differ == 3)TRUE                           0.441082   0.213405   2.067   0.0387 *
I(rx == "Lev+5FU")TRUE:I(differ == 2)TRUE    0.321246   0.395361   0.813   0.4165
I(rx == "Lev+5FU")TRUE:I(differ == 3)TRUE    0.462637   0.438482   1.055   0.2914
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1839.0  on 905  degrees of freedom
Residual deviance: 1789.4  on 898  degrees of freedom
  (23 observations deleted due to missingness)
AIC: 2721.4

Number of Fisher Scoring iterations: 7
```

(g) We now use `anova()` to compare the two models. Explain how to interpret the test output. What can we conclude from these results in terms of the effects of treatment and tumour differentiation? (2 pts)

```
anova(fit,fit2,test="Chisq")

Analysis of Deviance Table

Model 1: status ~ I(age - 60) + I(rx == "Lev") + I(rx == "Lev+5FU") +
    I(differ == 2) + I(differ == 3) + offset(log(time/365.25))
Model 2: status ~ I(age - 60) + I(rx == "Lev") + I(rx == "Lev+5FU") +
    I(differ == 2) + I(differ == 3) + I(rx == "Lev+5FU"):I(differ ==
    2) + I(rx == "Lev+5FU"):I(differ == 3) + offset(log(time/365.25))
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       900     1790.5
2       898     1789.4  2   1.1589   0.5602
```

Part 2

Q2

(a) Explain why the analyses in Part 1 may **not** be suitable for investigating the time to recurrence? (1 pt)

Q3

We now use the `tamoxifen` dataset from the survival analysis textbook by Collett (2023). The dataset records results from a randomised controlled trial of tamoxifen use among breast cancer patients. We have the following variables:

`treat` randomised treatment arm (0=tamoxifen+radiation, 1=tamoxifen alone)

`psurv` progression-free survival time in days (no local, axillary or distant relapse, no second malignancy and no death)

`ps` event indicator for progression (0=No progression, 1=Progression)

`tsurv` survival time in days for any cause of death or last follow-up

`ts` event indicator for any cause of death (0=No, 1=Yes)

`age` age at study entry (years)

`size` tumour size (cm)

Notably, the coding for `treat` is binary 0/1 and where 1 is for tamoxifen alone, which is associated with less intensive treatment.

(a) The Kaplan-Meier estimators for progression-free survival and for overall survival are shown in Figure 1. Carefully describe and interpret the two sets of survival curves. (2 pts)

```
## Colour-blind palette of colours
cbPalette <- c("#999999", "#E69F00")
par(mfrow=1:2)
survfit(Surv(psurv/365.25, ps)~treat, data=tamoxifen) |>
    plot(xlab="Time since randomisation (years)",
```

```
        ylab="Survival",
        col=cbPalette[1:2], lwd = c(1.5,2), ylim=c(0.5,1), main="Progression-free survival")
survfit(Surv(tsurv/365.25, ts)~treat, data=tamoxifen) |>
    plot(xlab="Time since randomisation (years)", ylab="Survival",
        col=cbPalette[1:2], lwd = c(1.5,2), ylim=c(0.5,1), main="All causes of death")
legend("bottomleft", legend=c("treatment: tamoxifen+radiotherapy","treatment: tamoxifen"),
       col=cbPalette[1:2], lwd=c(1.5,2), lty=1, bty="n")
```

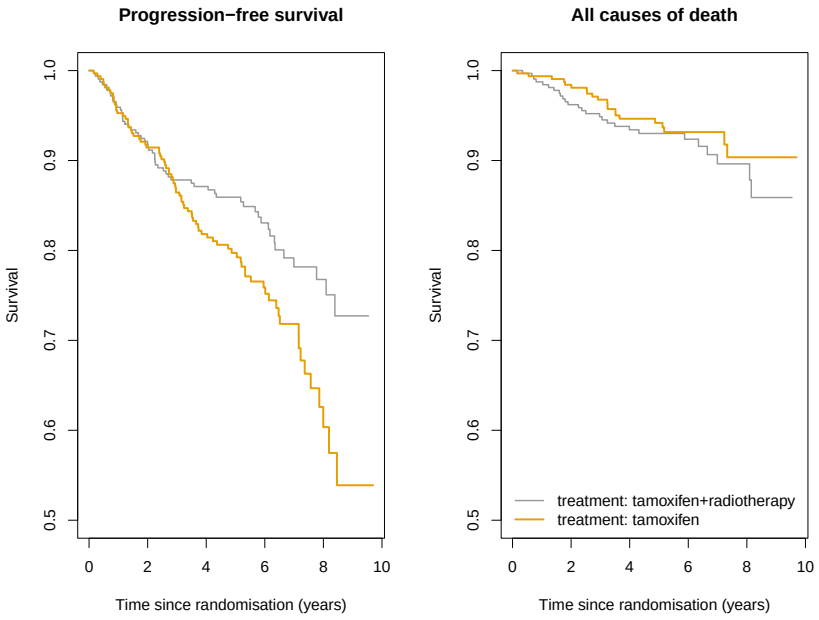

Figure 1: Kaplan-Meier survival curves for progression-free survival and for all causes of death by randomised treatment assignment

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

```
fit = coxph(Surv(psurv,ps)~treat+I(size>=2), data=tamoxifen)
summary(fit)

Call:
coxph(formula = Surv(psurv, ps) ~ treat + I(size >= 2), data = tamoxifen)

  n= 641, number of events= 138

                   coef exp(coef) se(coef)     z Pr(>|z|)
treat            0.4328    1.5415   0.1735 2.494   0.0126 *
I(size >= 2)TRUE 0.7908    2.2052   0.1706 4.637 3.54e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                 exp(coef) exp(-coef) lower .95 upper .95
treat                1.542     0.6487     1.097     2.166
I(size >= 2)TRUE     2.205     0.4535     1.579     3.081

Concordance= 0.62  (se = 0.026 )
Likelihood ratio test= 26.65  on 2 df,   p=2e-06
Wald test            = 27.26  on 2 df,   p=1e-06
Score (logrank) test = 28.44  on 2 df,   p=7e-07
```

(c) Based on the previous output, discuss whether there is any evidence that treatment for tamoxifen alone is associated with progression-free survival after adjusting for tumour size. Provide confidence intervals and p-values to support your argument. (2 pts)

(d) We are interested in whether the effect of treatment on progression varies by tumour size. We fit a Cox model that includes main effects for treatment, a main effect for `I(size>=2)`, and interactions between treatment and `I(size>=2)`. Based on the regression model output, summarise and discuss the evidence for whether the treatment effect on progression varies by tumour size. (2 pts)

```
coxph(Surv(psurv,ps)~treat*I(size>=2), data=tamoxifen) |>
    summary()

Call:
coxph(formula = Surv(psurv, ps) ~ treat * I(size >= 2), data = tamoxifen)

  n= 641, number of events= 138

                         coef exp(coef) se(coef)     z Pr(>|z|)
treat                  0.2667    1.3056   0.2373 1.124   0.2611
I(size >= 2)TRUE       0.5841    1.7934   0.2671 2.187   0.0287 *
treat:I(size >= 2)TRUE 0.3517    1.4215   0.3478 1.011   0.3119
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                       exp(coef) exp(-coef) lower .95 upper .95
treat                      1.306     0.7659     0.820     2.079
I(size >= 2)TRUE           1.793     0.5576     1.063     3.027
```

```
treat:I(size >= 2)TRUE     1.421     0.7035     0.719     2.810

Concordance= 0.62  (se = 0.026 )
Likelihood ratio test= 27.67  on 3 df,   p=4e-06
Wald test            = 30.22  on 3 df,   p=1e-06
Score (logrank) test = 32.87  on 3 df,   p=3e-07
```

(e) To assess non-proportionality, we can use Schoenfeld residuals from a Cox regression model to (i) test for non-proportionality and (ii) plot for a smoothed log hazard ratio. See the table and plot. Carefully interpret the findings. (4 pts)

```
cox.zph(fit)
```

```
             chisq df     p
treat         5.24  1 0.022
I(size >= 2)  1.69  1 0.194
GLOBAL        6.64  2 0.036
```

```
plot(cox.zph(fit), var="treat")
```

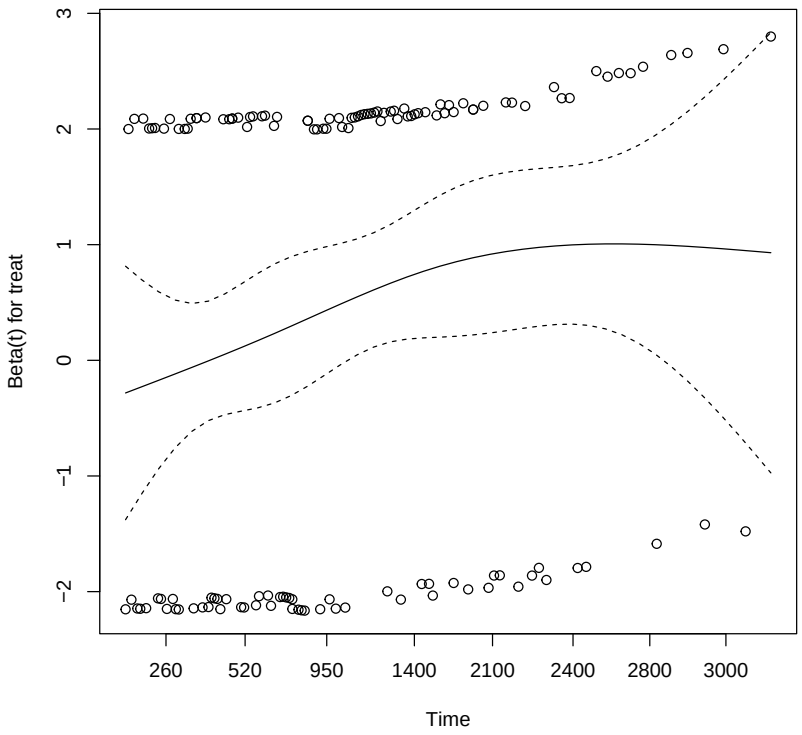

Figure 2: Schoenfeld residual plot for the association between progression-free survival and treatment by time since diagnosed

(f) We can model for a time-varying hazard ratio using the `tt` argument in `coxph`. Write out a formula for the modelled hazard. For a given size, what is the hazard ratio for treatment at 0 and 1 years? (3 pts)

```
coxph(Surv(psurv,ps)~treat+I(size>=2)+tt(treat),
      data=tamoxifen, tt=function(x,t,...) x*t) |> summary()

Call:
coxph(formula = Surv(psurv, ps) ~ treat + I(size >= 2) + tt(treat),
    data = tamoxifen, tt = function(x, t, ...) x * t)

  n= 641, number of events= 138

                       coef  exp(coef)   se(coef)      z Pr(>|z|)
treat             -0.0682921  0.9339877  0.2951471 -0.231   0.8170
I(size >= 2)TRUE  0.8082307  2.2439343  0.1707760  4.733 2.22e-06 ***
tt(treat)         0.0004347  1.0004348  0.0002115  2.055   0.0399 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                  exp(coef) exp(-coef) lower .95 upper .95
treat                 0.934     1.0707    0.5237     1.666
I(size >= 2)TRUE      2.244     0.4456    1.6056     3.136
tt(treat)             1.000     0.9996    1.0000     1.001

Concordance= 0.625  (se = 0.026 )
Likelihood ratio test= 30.98  on 3 df,   p=9e-07
Wald test            = 31.23  on 3 df,   p=8e-07
Score (logrank) test = 32.87  on 3 df,   p=3e-07
```

Q4

(a) Drawing on your own research (or from the course material), select a time to event of interest Y, with an exposure variable X and another covariate U. Write a Methods section for an article describing an analysis for whether the event of interest Y is related to exposure X, possibly adjusting for, or interacting with, covariate U. The Methods should include: the general study design, including study inclusion and exclusion criteria; how Y, X and U are measured; which estimands are being considered; which models and estimators are used; and any other statistical methods. You will be judged on novelty and completeness of your reporting. (5 pts)

(b) Hernán (2010; `https://doi.org/10.1097/EDE.0b013e3181c1ea43`) cautions about the use of hazard ratios in epidemiology. Based on the article and the course material, which estimands should we consider using to compare time-to-event for two groups adjusting for potential confounders? (3 pts)

(Part 1: 17 pts; Part 2: 24 pts)