BIOSTAT III: Survival Analysis for Epidemiologists in R

Take-home examination

11–20 November, 2019

Instructions

- **The examination is individual-based: you are not allowed to cooperate with anyone**, although you are encouraged to consult the available literature. The teachers will use Urkund in order to assess potential plagiarism (`http://ki.se/sites/default/files/cheating_is_forbidden_2013.pdf`)
- The examination will be made available by 12:00 on Wednesday 20 November 2019 and **the examination is due by 17:00 on Wednesday 27 November 2019**. Please contact Mark Clements before the due date and time if you wish to request an extension.
- The examination will be graded and results returned to you by Friday 6 December 2019.
- The examination is in two parts. You need to score at least 7/13 for Part 1 and 14/25 in Part 2 to pass the examination.
- The examination dataset is available from `http://biostat3.net/download/exams/2019_R/`.
- Do not write answers by hand: please use Word, LaTeX or a similar format for your examination report.
- Motivate all answers in your examination report, but write an answer that is as brief as possible without loss of clarity. Define any notation that you use for equations. The examination report should be written in English.
- Provide key computer output within the text.
- **You are expected to write computer code to read and analyse the data.** Include your computer code in your report. You are encouraged to use R, Stata or SAS for your analysis; if you wish to use other software, please contact Mark Clements (`mark.clements@ki.se`).
- 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.**

Description of simulated data for prostate cancer testing

Both parts of the exam use simulated data for a prostate cancer screening trial. The prostate is a male reproductive organ responsible for producing semen. In men aged over age 60 years, there is a high likelihood that a man has a prostate has cancer but with no symptoms. For many men with prostate cancer, the disease progresses very slowly and many men will die due to other causes without any symptoms due to the cancer. However, for some men with prostate cancer, the cancer will progress more quickly leading to symptoms and possibly prostate cancer death. In Sweden, prostate cancer accounts for a third of all male cancer diagnoses and is the leading cause of male cancer death.

We simulated for a randomised controlled prostate cancer screening trial with two trial arms: (1) no screening between ages 50 and 75 years; and (2) two-yearly screening between ages 50-69 years with follow-up to age 75 years. Each arm had 96,607 men with no diagnosis of prostate cancer before age 50 years followed from age 50 years through to age 75 years or death, whichever happened first.

For the unscreened arm, men were diagnosed clinically with prostate cancer following symptoms and treated with standard clinical care. For the screening arm, men had a prostate-specific antigen (PSA) test every two years; for men with a PSA test over 3 ng/mL, the men were referred to a urologist for a biopsy, with 85% biopsy compliance. Men in the screening arm followed the same clinical care as per the unscreened arm.

Data were recorded for age at study entry (50 years), PSA at study entry, age at prostate cancer diagnosis (if any), age at the end of follow-up, and the man's status at the end of follow-up, including a possible cause of death or whether censored.

You have been provided an analysis dataset in the examination folder. The dataset is called `prostate.csv`, which is a comma-separated values (text) file. You should read the .csv file into your statistical software:

R:

```
prostate <- read.csv("http://biostat3.net/download/exams/2019_R/prostate.csv")
```

Stata:

```
import delimited "http://biostat3.net/download/exams/2019_R/prostate.csv", clear
```

SAS:

```
filename afile url "http://biostat3.net/download/exams/2019_R/prostate.csv";
data prostate;
    infile afile delimiter="," dsd firstobs=2;
    input id screening age_start psa_start age_dx event_dx age_dth event_dth;
run;
* or download the file locally and...;
proc import datafile="prostate.csv" out=prostate replace;
run;
```

The columns for the `prostate.csv` file are:

Variable name	Description	Encoding
id	Individual identification number	Integer, between 1 and 200000
screening	Trial arm	1 = screening arm 0 = unscreened arm
age_start	Age at study entry (years)	Float, 50.0
psa_start	PSA value at study entry (ng/mL)	Float, >0
age_dx	Age for prostate cancer diagnosis (years)	Float, 50–75 years
event_dx	Event indicator for prostate cancer diagnosis	1=diagnosed, 0=censored
age_dth	Age for death outcomes (years)	Float
event_dth	Event indicator for death	Integer, 0=censored, 1=prostate cancer death, 2=other cause of death

For Part 1, we also provide collapsed data for prostate cancer mortality from study entry. The dataset is called `mortality.csv`, which is a comma-separated values (text) file. You can read the .csv file into your statistical software using:

R:

```
mortality <- read.csv("http://biostat3.net/download/exams/2019_R/mortality.csv")
```

Stata:

```
import delimited "http://biostat3.net/download/exams/2019_R/mortality.csv", clear
```

SAS:

```
filename afile url "http://biostat3.net/download/exams/2019_R/mortality.csv";
data mortality;
    infile afile delimiter="," dsd firstobs=2;
    input screening age pt n;
run;
* or download the file locally and...;
proc import datafile="mortality.csv" out=collapsed replace;
run;
```

The columns for the `mortality.csv` file are:

Variable name	Description	Encoding
screening	Trial arm	1 = screening arm, 0 = unscreened arm
age	Attained age (years)	Integer: 50, 51, . . . , 74
pt	Person-time from study entry to prostate cancer death or censoring	Float, person-years
n	Number of prostate cancer deaths	Integer, ≥ 0

Part 1

In Part 1, you can use either collapsed data or the unit record data (which will need to be split). For splitting the unit record data, you could use the following:

```
prostate2 <- survSplit(Surv(age_start,age_dth,event_dth==1)~., data=prostate, cut=50:75)
prostate2 <- transform(prostate2, n=event, pt=age_dth-age_start, age=age_start)
```

Question 1

Carefully describe the prostate cancer mortality rate pattern in Figure 1. (1 pt)

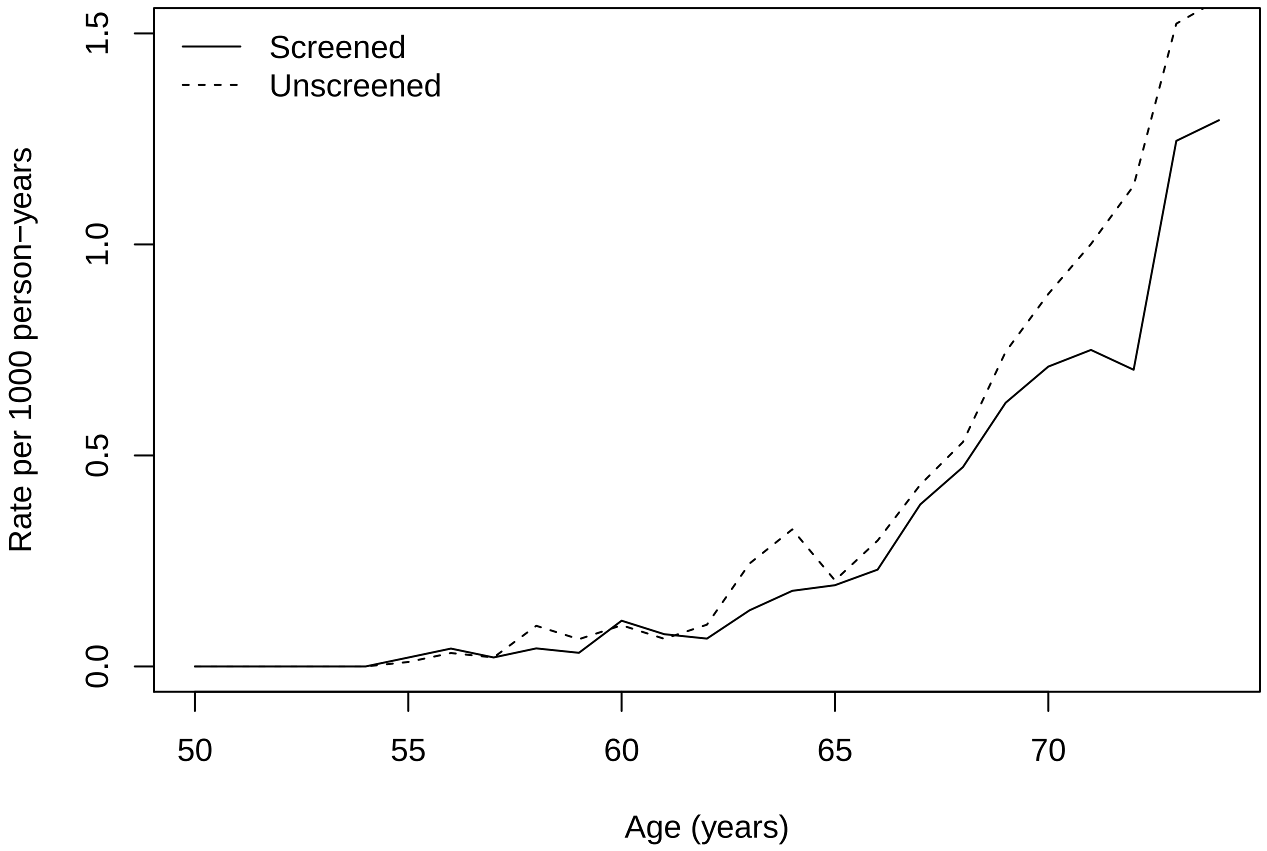

Figure 1: Age-specific prostate cancer mortality rates by screening arm

Question 2

(a) Estimate the prostate cancer mortality rate ratio and 95% confidence interval comparing the screening arm with the unscreened arm for ages 50–74 years. As a reminder, describe your analytical approach, show your code and output, and interpret your findings. (2 pts)

(b) Write out a formula for a Poisson regression model to estimate a prostate cancer mortality rate ratio comparing the screening arm with the unscreened arm, adjusting for age as a linear effect. *(Reminder: please explain your notation.)* (2 pts)

(c) Using the model from (b), estimate the prostate cancer mortality rate ratio and 95% confidence interval comparing the screening arm with the unscreened arm. (2 pts)

(d) Separately for ages 60–64, 65–69 and 70–74 years, estimate the prostate cancer mortality rate ratios and 95% confidence intervals comparing the screening arm with the unscreened arm. (2 pts)

(e) Write out a formula for a Poisson regression model to compare the three prostate cancer mortality rate ratios in (d), adjusting for age as a linear effect. (2 pts)

(f) Fit the model in (e) and interpret whether there is evidence that the three prostate cancer mortality rate ratios are different. (2 pts)

Part 2

Question 3

In the following question, consider the following *two* outcomes: (i) prostate cancer incidence from study entry; and (ii) prostate cancer mortality from study entry.

(a) Time since study entry is one possible *time scale* (that is, time origin and unit of time) of interest for these outcomes. Discuss other time scales and their advantages or disadvantages for both outcomes. (2 pts)

(b) For each of the two outcomes, plot and carefully interpret the Kaplan-Meier curves by trial arm using time since study entry as the time scale. (2 pts)

(c) For each outcome and using time since study entry as the time scale, use Cox regression to estimate the (time-constant) hazard ratio and 95% confidence interval for the screening trial arm compared with the unscreened trial arm for ages 50–74 years. (2 pts)

(d) For prostate cancer mortality, compare the hazard ratio for the screening arm compared with the unscreened arm from (c) with the rate ratios from 2(a) and 2(c). Why are these three estimates not identical? (2 pts)

(e) Summarise your findings for the two outcomes. Also discuss any potential biases. (3 pts)

Question 4

In the following question, restrict to men diagnosed with prostate cancer between age 50 and 75 years and consider the time from prostate cancer diagnosis to prostate cancer death (or study exit).

(a) Using the available data, describe those diagnosed with prostate cancer by trial arm (e.g. number, median age, median PSA at age 50 years). (2 pt)

(b) Fit a proportional hazards flexible parametric survival model for cause-specific survival from diagnosis (that is, prostate cancer mortality), using time from diagnosis as the time scale. Is there any evidence that survival is different in the screened arm compared with the unscreened arm? (2 pts)

(c) Specify, fit and interpret a regression model to estimate time-varying hazard ratios for cause-specific survival for the screened versus the unscreened arm. (3 pts)

(d) Summarise your findings for cause-specific survival. (2 pts)

Question 5

(a) Explain the difference between *left censoring* and *left truncation*. (2 pts)

(b) For time-to-event data, we can define a *random effect* as a covariate Z (or where $\exp(Z)$ is a *frailty*) that is associated with the time to event but not associated with the exposure X of interest. For a proportional hazards model, what are two sufficient conditions when we would estimate similar hazard ratios for the exposure of interest X when we either do or do not adjust for Z? (3 pts)