BIOSTAT III: Survival Analysis for Epidemiologists in R

Take-home examination

5–14 November, 2018

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 17:00 on Wednesday 14 November 2018 and **the examination is due by 17:00 on Monday 26 November 2018**. 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 Monday 3 December 2018.
- The examination is in two parts. You need to score at least 7/13 for Part 1 and 11/21 in Part 2 to pass the examination.
- The examination dataset is available from `http://biostat3.net/download/exams/2018_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/2018_R/prostate.csv")
```

Stata:

```
import delimited "http://biostat3.net/download/exams/2018_R/prostate.csv", clear
```

SAS:

```
filename afile url "http://biostat3.net/download/exams/2018_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 incidence. The dataset is called `collapsed.csv`, which is a comma-separated values (text) file. You can read the .csv file into your statistical software using:

R:

```
collapsed <- read.csv("http://biostat3.net/download/exams/2018_R/collapsed.csv")
```

Stata:

```
import delimited "http://biostat3.net/download/exams/2018_R/collapsed.csv", clear
```

SAS:

```
filename afile url "http://biostat3.net/download/exams/2018_R/collapsed.csv";
data collapsed;
    infile afile delimiter="," dsd firstobs=2;
    input screening age pt n;
run;
* or download the file locally and...;
proc import datafile="collapsed.csv" out=collapsed replace;
run;
```

The columns for the `collapsed.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 diagnosis or censoring	Float, person-years
`n`	Number of prostate cancer diagnoses	Integer, ≥ 0

Part 1

In Part 1, you can use either the unit record data or the collapsed data.

Question 1

Plot the prostate cancer incidence rates by single-year age groups and trial arm. Carefully describe the rate patterns. (3 pts)

Question 2

(a) Estimate the incidence rate ratios and 95% confidence intervals comparing the screening arm with the unscreened arm for (i) ages 50-69 years and (ii) ages 70-74 years. As a reminder, describe your analytical approach, show your code and output, and interpret your findings. (2 pts)

Formally compare the two rate ratios using a regression model:

(b) Write out a formula for the regression model to compare the two rate ratios. *(Reminder: please explain your notation.)* (2 pts)

(c) Fit the model in (b) and interpret whether there is evidence that the two rate ratios are different. (2 pts)

(d) Is there any evidence that the rate ratio comparing the screening arm with the unscreened arm varies by age between ages 50 and 69 years? Specify, fit and report on one or more models to address this question. (2 pts)

(e) Discuss whether you need, and how, to adjust for potential confounding variables in Question 2. (2 pts)

Part 2

Question 3

In the following question, consider the following *four* outcomes: (i) prostate cancer incidence from study entry; (ii) prostate cancer mortality from study entry; (iii) all cause mortality from study entry; and (iv) prostate cancer cause-specific survival from prostate cancer diagnosis.

(a) Time since study entry is one possible *time scale* of interest for these outcomes. Discuss other time scales and their advantages or disadvantages for each of the four outcomes. (2 pts)

(b) For each of the four outcomes, choose a primary time scale, plot and carefully interpret the Kaplan-Meier curves by trial arm. (4 pts)

Question 4

(a) For each of the four outcomes and using the time scale in 3(b), 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. (4 pts)

(b) For each of the four outcomes, specify, fit and interpret regression models to estimate time-varying hazard ratios between ages 50 and 75 years. For each outcome, do we expect the hazard ratio to be smooth or discontinuous? (4 pts)

(c) Summarise your findings for the four outcomes. Also discuss any potential biases. (4 pts)

Question 5

Explain the concept of risk sets for Cox regression. Explain the relationship between a cohort study and a nested case-control study in terms of risk sets. Using this relationship, discuss for a nested case-control study whether a control can be a control for more than one case and whether a case can become a control. (3 pts)