BIOSTAT III: Survival Analysis for Epidemiologists in Stata

Take-home examination

11–20 February, 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 examiner will use Urkund in order to assess potential plagiarism (`http://ki.se/sites/default/files/cheating_is_forbidden_2013.pdf`) and use similarities in code to assess unauthorised cooperation.
- The examination will be made available by 17:00 on Wednesday 20 February 2010 and **the examination is due by 17:00 on Wednesday 27 February 2019**.
- The examination will be graded and results returned to you by Wednesday 6 March 2019.
- The examination is in two parts. You need to score at least 7/12 for Part 1 and 12/22 in Part 2 to pass the examination.
- The examination dataset is available from `http://biostat3.net/download/exams/2019/`.
- Do not write answers by hand: please use Word, LaTeX or a similar format for your examination report and provide 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.
- 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 analysis dataset in two formats in the examination folder. The dataset `prostate.csv` is a comma-separated values (text) file. The dataset `prostate.dta` is a Stata file that is compatible with Stata 11 or later. You should read the .dta or .csv file into your statistical software:

Stata:

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

R:

```
library(foreign)
prostate <- read.dta("http://biostat3.net/download/exams/2019/prostate.dta")
## or
prostate <- read.csv("http://biostat3.net/download/exams/2019/prostate.csv")
```

SAS:

```
filename afile url "http://biostat3.net/download/exams/2019/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)	Continuous/float, 50.0
`psa_start`	PSA value at study entry (ng/mL)	Continuous/float, >0
`age_dx`	Age for prostate cancer diagnosis (years)	Continuous/float, 50–75 years
`event_dx`	Event indicator for prostate cancer diagnosis	1=diagnosed, 0=censored
`age_dth`	Age for death outcomes (years)	Continuous/float
`event_dth`	Event indicator for death	Integer, 0=censored, 1=prostate cancer death, 2=other cause of death

Part 1

In Part 1, we will assume that rates are constant on all time scales.

Question 1

(a) Estimate the average prostate cancer incidence rates and 95% confidence intervals comparing the screening arm with the unscreened arm. *Reminder, describe your analytical approach, show your code and output, and interpret your findings.* (2 pts)

(b) Using Poisson regression, estimate and interpret the prostate cancer incidence rate ratio and 95% confidence intervals comparing the screening arm with the unscreened arm. (2 pts)

(c) Discuss whether you need, and how, to adjust for potential confounding variables in Question 1. (2 pts)

Question 2

For this question, restrict to men with an initial PSA value of 3 ng/mL or over.

(a) Estimate the prostate cancer incidence rate ratio and 95% confidence intervals comparing the screening arm with the unscreened arm for (i) men with a 3 $\leq$PSA$<$ 10 ng/mL and (ii) for men with a PSA of 10 ng/mL or over. (2 pts)

(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)

Part 2

In Part 2, consider the following *three* outcomes: (i) death due to prostate cancer among men diagnosed with prostate cancer; (ii) death due to any cause for men diagnosed with prostate cancer; and (iii) prostate cancer incidence from study entry.

Question 3

(a) For survival from prostate cancer diagnosis (outcomes (i) and (ii)), time since prostate cancer diagnosis is one possible *time scale* of interest. Discuss other time scales and their advantages or disadvantages for cause-specific survival and all cause survival. (2 pts)

(b) For both of these outcomes, choose a primary time scale, plot survival and hazards and carefully interpret the results by trial arm. (4 pts)

Question 4

For this question, we will compare Cox regression with Poisson regression in Question 1 for outcome (iii).

(a) For time from study entry to prostate cancer incidence (outcome (iii)), choose a primary time scale, fit a Cox regression model to compare by trial arm using a time-constant hazard ratio, and interpret the hazard ratio. (2 pts)

(b) Compare the estimated hazard ratio in Question 4(a) with the incidence rate ratio in Question 1(b). Discuss why these estimates are *not* the same. (2 pts)

Question 5

(a) For time from study entry to prostate cancer incidence (outcome (iii)), use Schoenfeld residuals to investigate whether there is evidence for non-proportional hazard ratios by trial arm (i.e. for the hazard ratio comparing the screening arm with the unscreened arm). (3 pts)

(b) Using either Poisson regression, Cox regression or flexible parametric models, estimate a time-varying hazard ratio for outcome (iii) by trial arm. (3 pts)

(c) Summarise your findings for the trial by outcomes (i)–(iii). (2 pts)

Question 6

Discuss the similarities and differences in (a) study design and (b) analysis methods for: (i) cohort studies analysed using Cox's proportional hazards regression model; (ii) nested case-control studies; and (iii) case-cohort studies. (4 pts)