BIOSTAT III: Survival Analysis for Epidemiologists

Take-home Examination

15–27 March, 2017

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 at 12:00 on Wednesday 15 March 2017 and the examination is due by 17:00 on Monday 27 March 2017.
- The examination will be graded and results will be returned to you by 7 April 2017.
- The examination is in two parts. You need to score at least 5/9 for Part 1 and 11/20 in Part 2 to pass the examination.
- Students who do not obtain a passing grade in the first examination will be offered a second examination within 2 months of the final day of the course.
- The examination dataset is available from `http://biostat3.net/download/exams/2017/`.
- 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. Provide your computer code as an appendix.
- **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 Stata, R 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 in the document containing the answers.**

Description of simulated data for prostate cancer testing

Both parts of the exam use simulated data for prostate cancer testing in Stockholm. The prostate is a male reproductive organ responsible for producing semen. In older men, there is a high likelihood that the 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.

Prostate cancer testing using the prostate-specific antigen (PSA) test is very common in Stockholm. The PSA test is a simple and inexpensive blood test which returns a value measured in terms of ng/mL. Most men have a PSA value between 0 and 3 ng/mL; for such values, few men will be referred to a urologist. For men with a PSA value that is 3 ng/mL or over, men are more often referred to a urologist for a prostate biopsy which may lead to a prostate cancer diagnosis. The risk of prostate cancer increases with higher PSA values, with a PSA of 10 ng/mL or over associated with a high risk of prostate cancer. Despite the high levels of PSA testing, PSA testing is *not* organised and there is marked uncertainty in whether the benefits from PSA testing out-weigh the potential harms from over-diagnosis and over-treatment.

For the simulated data, men enter the cohort at their initial PSA test during the period 2003–2015 with no previous prostate cancer diagnosis. They are then followed for the following outcomes: (i) prostate cancer incidence; and (ii) death, either due to prostate cancer or other causes of death. Censoring may occur due to emigration, death due to another cause or end of follow-up at 31 December 2015. The research questions relate to the risk of a prostate cancer diagnosis from their initial PSA test, or for the risk of prostate cancer death from their initial PSA test. Covariates include age and the PSA values.

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

Stata:

```
import delimited "http://biostat3.net/download/exams/2017/psa.csv", clear
```

SAS:

```
filename afile url "http://biostat3.net/download/exams/2017/psa.csv";
data psa;
    infile afile delimiter="," dsd firstobs=2;
    input id psa start_age age_dx event_dx age_dth event_dth;
run;
* or download the file locally and...;
proc import datafile="psa.csv" out=psa replace;
run;
```

R:

```
psa <- read.csv("http://biostat3.net/download/exams/2017/psa.csv")
```

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

Variable name	Description	Encoding
`id`	Individual identification number	Integer, 1 to 100000
`psa`	PSA value (ng/mL)	Float, 0.005 to 17585.3
`start_age`	Age at initial PSA test (years)	Float, 50 to 79
`age_dx`	Age for prostate cancer diagnosis outcome (years)	Float
`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

Part 1

For this part, we are interested in the time from the initial PSA test to prostate cancer diagnosis.

Question 1

We initially describe the cohort at study entry for the initial PSA test. Create PSA categories with cut-points of 0, 1, 2, 3, 10 and 17586 ng/mL. Create ten-year age groups (50–59, 60–69 and 70-79 years).

Tabulate and interpret how the proportions for the PSA categories vary by age groups. Formally test and interpret whether age is associated with PSA values. [2pt]

Question 2

To simplify the Poisson regression analyses, *restrict the analysis to men aged 50–69 years at their initial PSA test with PSA values less than 3 ng/mL*. Assume that the rates are constant over the time scale. The focus is on modelling rates for lower PSA values.

(a) Using Poisson regression, model the rates for prostate cancer incidence with main effects for age groups and PSA categories. Report rate ratios, 95% confidence intervals and p-values, and interpret the associations [2pt]

(b) Propose, fit and interpret a model for prostate cancer incidence rates to assess whether there is an interaction between age groups and PSA categories. [2pt]

(c) Write out a formula for the regression model in (b). (Reminder: explain your notation.) [1pt]

(d) Using the model from (b), calculate the prostate cancer incidence rate and 95% confidence intervals for a man aged 62 years with a PSA value of 1.1 ng/mL. [1pt]

(e) Using the rate and 95% confidence interval from (d), calculate the ten-year risk of prostate cancer incidence and 95% confidence intervals for a man aged 62 years with a PSA value of 1.1 ng/mL. [1pt]

Part 2

For this part, we are interested in the time from the initial PSA test to prostate cancer death, censoring for other causes of death.

Question 3

To simplify the following analysis, *restrict to men aged 60–69 years at their initial PSA test.*

(a) Plot and interpret the Kaplan-Meier survival curves for prostate cancer death by PSA categories. [2pt]

(b) Formally assess and interpret whether survival for prostate cancer death varies by PSA categories. [1pt]

(c) Estimate the ten-year risks and 95% confidence intervals for prostate cancer death by PSA categories. [1pt]

Question 4

To simplify the Cox regression analyses, *restrict the analysis to men aged 50–69 years at their initial PSA test with PSA values less than 3 ng/mL.* The focus is on modelling mortality for lower PSA values.

(a) Using Cox regression, fit and interpret a model for prostate cancer mortality rates adjusting for age groups and PSA categories. [2pt]

(b) Discuss the choice of time scale for the model in (a). [1pt]

(c) Write out a formula for the regression model in (a). (Reminder: explain your notation.) [1pt]

(d) Provide a mathematical formula for how to predict ten-year risk for a man aged 62 years with a PSA value of 1.1 ng/mL from baseline survival $S_0(10)$ and hazard ratio $HR(age = 62, PSA = 1.1)$. [1pt]

(e) Contrast the assumptions underlying risk estimates in 3 (c) and 4 (d). [1pt]

(f) Assess whether the hazard ratios for PSA categories varies by time since the initial PSA test. Use *one* different approach to test for and estimate time-dependent hazard ratios, using either (i) piece-wise constant hazard ratios with time splitting, (ii) a continuous time-varying effect (e.g. using the `tvc` and `texp` options in Stata), or (iii) a flexible parametric survival model (e.g. `stpm2`). Interpret the output. [2pt]

(g) Fit a stratified Cox model that is stratified by PSA category and adjusted for age category. Report the hazard ratios and 95% confidence intervals. Contrast and discuss the model assumptions for 4 (a) and 4 (g). [2pt]

Question 5

(a) Summarise the findings from Questions 1–4 and discuss whether it would be safe for men with PSA values below 3 ng/mL to not be tested for 5 or 10 years. You can supplement the above analyses with further analyses. [3pt]

(b) How would the prostate cancer incidence and mortality risk estimates have changed if PSA testing had been uncommon? How does that affect your conclusions in (a)? [1pt]

Question 6

What is the mathematical equation for Cox's partial likelihood and for nested case-control studies? How are they different? Discuss the interpretation of the odds ratio from a nested case-control study. [2pt]