BIOSTAT III: Survival Analysis for Epidemiologists in R

Take-home examination

21–29 November, 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 by 17:00 on Tuesday 21 November 2017 and **the examination is due by 17:00 on Wednesday 29 November 2017**.
- The examination will be graded and results will be returned to you by Monday 11 December 2017.
- The examination is in two parts. You need to score at least 6/9 for Part 1 and 8/15 in Part 2 to pass the examination.
- The examination dataset is available from `http://biostat3.net/download/exams/2017_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 in the document containing the answers.**

Part 1

Description of simulated lung cancer incidence

Lung cancer is a common cancer in many countries, with high incidence rates, largely attributable to smoking exposure, with poor survival and high mortality rates. In the following analysis, we consider possible causes of lung cancer incidence, including smoking and asbestos exposure, and potential confounding factors, including sex and attained age.

You have been provided *collapsed data* for analysis in the re-examination folder. For these data (that is, numbers and person-time which can be used for rate calculations). Note that you cannot use the `Surv` function for Part 1. The dataset is called `incidence.csv`, which is a comma-separated values (text) file. You should read the .csv file into your statistical software:

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

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

SAS:
```
filename afile url "http://biostat3.net/download/exams/2017_R/incidence.csv";
data incidence;
    infile afile delimiter="," dsd firstobs=2;
    input sex smoking asbestos age pt lc;
run;
* or download the correct file locally and...;
proc import datafile="incidence.csv" out=incidence replace;
run;
```

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

Variable name	Description	Encoding
`sex`	Sex	1=Males, 0=Females
`smoking`	Life-time exposure to cigarette smoking	1=Current, 0=Never
`asbestos`	Asbestos exposure	1=Exposed, 0=Unexposed
`age`	Age of follow-up	Single year of age
`pt`	Aggregated person-time of follow-up	Person-years
`lc`	Total number of incident `lung cancer` cases	Number

Question 1

To simplify this question, ignore ages 80 years and over. Recode age into 10-year age groups (40–49, 50–59, 60–69 and 70–79 years). Report the lung cancer incidence rates by

age and sex with 95% confidence intervals and describe the pattern. [2pt]

Question 2

Investigate the association between lung cancer incidence and sex.

(a) Without adjusting for other variables, report the rate ratio comparing lung cancer incidence rates in males compared with females, with the 95% confidence interval and the p-value. [1pt]

(b) Using Poisson regression, fit a main effects model with age, sex and smoking status. Report the adjusted rate ratios with 95% confidence intervals and p-values. Is there any evidence that the association between lung cancer and sex is confounded? *(Hint: do not forget to incorporate person-time into the Poisson regression.)* [2pt]

(c) For the fitted model in (b), write out a formula for the regression model. *(Reminder: please explain your notation.)* [1pt]

Question 3

(a) What model would you fit to assess whether the smoking rate ratios for males and females are different? Fit this model, report the sex-specific smoking rate ratios and interpret whether there is a difference. [3pt]

Part 2

Description of a simulated randomised trial for lung cancer survival

The incident lung cancer cases from Part 1 are assumed to be recruited to a randomised controlled trial of lung cancer treatment, comparing conventional therapy (chemotherapy) with a combination of chemotherapy and radiotherapy. The lung cancer patients are followed for up to five years.

The dataset is called `survival.csv`, which is a comma-separated values (text) file. You should read the .csv file into your statistical software:

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

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

SAS:
```
filename afile url "http://biostat3.net/download/exams/2017_R/survival.csv";
data survival;
    infile afile delimiter="," dsd firstobs=2;
    input id age sex asbestos smoking tx tsurv event;
run;
* or download the correct file locally and...;
```

```
proc import datafile="survival.csv" out=survival replace;
run;
```

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

Variable name	Description	Encoding
`id`	Row/individual ID	$1, \ldots, \#\text{rows}$
`age`	Age at cancer diagnosis	Years
`sex`	Sex	1=Males, 0=Females
`asbestos`	Asbestos exposure	1=Exposed, 0=Unexposed
`smoking`	Life-time exposure to cigarette smoking	1=Current, 0=Never
`tx`	Randomised treatment modality	0=Conventional (chemo.), 1=Chemo.+radio.
`tsurv`	Event time	Years from diagnosis
`event`	Status at end of follow-up	1=Lung cancer death, 0=Otherwise

Question 4

(a) Time since cancer diagnosis is one possible *time scale* of interest. Discuss other time scales and their advantages or disadvantages. [1pt]

(b) For lung cancer mortality as the outcome and time since diagnosis as the time scale, plot and interpret the Kaplan-Meier curves by sex and broad age groups. You should motivate your choice of age groups. [2pt]

(c) Perform and interpret a test to assess whether these curves are different. [1pt]

Question 5

(a) Using Cox regression, estimate the hazard ratio and 95% confidence interval for males compared with females, possibly adjusting for potential confounding covariates. Discuss any adjustment for potential confounding variables and interpret the hazard ratio. [2pt]

(b) Using Cox regression, estimate the hazard ratio and 95% confidence interval comparing chemotherapy+radiotherapy with conventional chemotherapy. Discuss any adjustment for potential confounding variables and interpret the hazard ratio. [2pt]

(c) For comparing lung cancer treatment modalities, how would model building for an observational study differ from model building for a randomised controlled trial? [2pt]

Question 6

Discuss how to model for effect modification between time and a covariate using Poisson regression, Cox regression and flexible parametric survival models. Your discussion should include (i) how to assess whether there is evidence for effect modification, (ii) how to estimate parameters for the effect modification, and (ii) the advantages and disadvantages of each of the three regression models for modelling effect modification. [5pt]