BIOSTAT III: Survival Analysis for Epidemiologists

Take-home Re-examination

16–27 May, 2016

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 09:00 on Monday 16 May 2016 and the examination is due by 17:00 on Friday 27 May 2016.
- The examination is in two parts. You need to score at least 5/8 for Part 1 and 4/7 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 datasets are available from `http://biostat3.net/download/exams/2016/`. You have been assigned a number for the specific folder below this web address for your examination datasets; you can find your number from Tab 2 of your printed material or by consultation with Mark Clements or Gunilla Nilsson Roos. **Specify your folder number in your examination report.**
- Do not write answers by hand: please use Word, LaTeX or a similar format for your examination report.
- Motivate all answers and show all calculations 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.
- **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.**

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, 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 your examination dataset folder (see Slide 200 of the lecture material). The dataset is called `incidence.csv`, which is a comma-separated values (text) file. You should read the .csv file into your statistical software (assuming folder number 1 – *note: change the folder number*):

Stata:

```
import delimited "http://biostat3.net/download/exams/2016/1/incidence.csv", clear
```

SAS:

```
filename afile url "http://biostat3.net/download/exams/2016/1/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;
```

R:

```
incidence <- read.csv("http://biostat3.net/download/exams/2016/1/incidence.csv")
```

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

What are the lung cancer incidence rates and 95% confidence intervals for males and females? [1pt]

Question 2

Using Poisson regression, investigate the association between lung cancer incidence and sex. We want to investigate whether any differences by sex are explained by other variables.

(a) Without adjustment for other variables, report the rate ratio, 95% confidence interval and p-value. [1pt]

(b) Using Poisson regression, investigate potential confounding with other variables. Motivate a final adjusted model and report the adjusted rate ratio for sex with its 95% confidence interval and p-value. [2pt]

(c) For the fitted model in (b), write out a formula for the regression model. [1pt]

Question 3

(a) Using separate analyses for males and females, and adjusting for age, what are the smoking rate ratios and 95% confidence intervals? [1pt]

(b) How would you test whether the smoking rate ratios for males and females are different? Perform this test and interpret the p-value. [2pt]

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 (assuming folder number 1 – *note: change the folder number*):

Stata:

```
import delimited "http://biostat3.net/download/exams/2016/1/survival.csv", clear
```

SAS:

```
filename afile url "http://biostat3.net/download/exams/2016/1/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;
```

R:

```
survival <- read.csv("http://biostat3.net/download/exams/2016/1/survival.csv")
```

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

Variable name	Description	Encoding
`id`	Row/individual ID	$1, \ldots, \#$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) For lung cancer mortality as the outcome and time since diagnosis as the time scale, plot and interpret the Kaplan-Meier curves by sex. [1pt]

(b) What is the five-year risk and 95% confidence interval of lung cancer death by sex? [1pt]

(c) How would you test for a difference between the two curves? Perform the test and interpret ther *p*-value. [1pt]

Question 5

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]

Question 6

Discuss whether the five-year risks of lung cancer death by sex could be affected by *competing risks*. How would the five-years risks change if there were no other causes of death? How would the risks change if there were no differences in other causes of death by sex? [2pt]