BIOSTAT III: Survival Analysis for Epidemiologists

Take-home Examination

16–28 March, 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 Wednesday 16 March 2016 and the examination is due by 17:00 on Monday 28 March 2016.
- The examination will be graded and results will be returned to you by 8 April 2016.
- The examination is in two parts. You need to score at least 8/15 for Part 1 and 9/17 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

Using Poisson regression, investigate the associations between lung cancer incidence rates and each of the following covariates: sex, asbestos exposure and smoking exposure. All of the analyses should be adjusted for age. For each association, report rate ratios, 95% confidence intervals and p-values, and interpret the association. [4pt]

Question 2

Using Poisson regression, fit a main effects model with age (linear effect), sex, smoking and asbestos exposure (Model A). For the fitted model, report the estimated rate ratios and their 95% confidence intervals for age, sex, smoking and asbestos exposure. Is there any empirical evidence for confounding between smoking and asbestos exposure? (Reminder: motivate your answers.) [4pt]

Question 3

Using Poisson regression, fit a model adjusted for age and sex to assess whether there is effect modification on a multiplicative scale between smoking and asbestos exposure (Model B).

(a) For this model, write out a formula for the regression model. (Reminder: explain your notation.) [2pt]

(b) Comparing Model B with Model A, is there evidence for effect modification on a multiplicative scale between smoking and asbestos exposure? (Reminder: motivate your answer.) [2pt]

(c) For the fitted model, what is the estimated lung cancer incidence rate (and 95% confidence interval) for males aged 62 years who were current smokers and who were exposed to asbestos? [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 (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, . . . , #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) Assess whether there is any statistical association between treatment modality and (i) age, (ii) sex, (iii) smoking exposure or (iv) asbestos exposure. [2pt]

(b) Plot and interpret the Kaplan-Meier curves by treatment modality. [2pt]

Question 5

Using Cox regression, estimate the hazard ratio and 95% confidence interval for chemotherapy+radiotherapy compared with conventional therapy (chemotherapy), possibly adjusting for potential confounding covariates. Discuss the choice of time scale and adjustment for potential confounding variables, and interpret the hazard ratio. [4pt]

Question 6

Assess whether the hazard ratio for treatment modality varies by time since diagnosis. Use *two* different approaches to test for and, where possible, estimate time-dependent hazard ratios, including either (i) an analysis using Schoenfeld residuals, (ii) using piecewise constant hazard ratios with time splitting, (iii) using a continuous time-varying effect (e.g. using the `tvc` and `texp` options in Stata), or (iv) a flexible parametric survival model (e.g. `stpm2`). Interpret the output. [6pt]

Question 7

(a) What are the advantages and disadvantages of using Poisson regression rather than Cox regression for addressing Questions 5–6? Which of Poisson regression or Cox regression would you use? Please motivate your choice. [2pt]

(b) Give a formula for a Poisson regression model to investigate whether the rate ratio for treatment modality is time-dependent. (Note: you do *not* need to fit this model.) [1pt]