BIOSTAT III: Survival Analysis for Epidemiologists

Take-home Examination

12–21 Feb, 2018

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 17:00 on Wednesday 21 Feb 2018 and the examination is due by 17:00 on Monday 26 Feb 2018. The examination will be graded and results will be returned to you by 5 March 2018.
- The examination is in two parts. You need to score at least 6/10 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.
- 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. The answers shoud be clear and concise. 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 colon cancer

Both parts of the exam use simulated data for colon cancer, where the primary research question is whether smoking has an effect on cancer-specific survival. The dataset is similar to the colon cancer data you have seen in the computer labs during the course. Below is a description of the data and the variables in the dataset. The examination dataset is available from `http://biostat3.net/download/exams/2018/`.

```
  obs:         4,623
 vars:            10
-------------------------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------------------------
id              int     %9.0g                 Unique patient ID
sex             byte    %8.0g      sex        Sex
age             byte    %9.0g                 Age at diagnosis
agecat          float   %9.0g      agec       Age at diagnosis, categorised
yydx            int     %9.0g                 Year of diagnosis
period          float   %14.0g     per        Year of diagnosis, categorised
stage           byte    %9.0g      stage      Clinical stage at diagnosis
smoke           float   %10.0g     smok       Smoking status
surv_mm         float   %9.0g                 Survival time in months (time since diagnosis)
status          byte    %17.0g     status     Vital status at exit
-------------------------------------------------------------------------------------------------

        Sex |      Freq.     Percent        Cum.
------------+-----------------------------------
     1=Male |      1,892       40.93       40.93
   2=Female |      2,731       59.07      100.00
------------+-----------------------------------
      Total |      4,623      100.00

     Age at |
 diagnosis, |
categorised |      Freq.     Percent        Cum.
------------+-----------------------------------
1=Age 16-49 |        396        8.57        8.57
2=Age 50-64 |      1,071       23.17       31.73
3=Age 65-74 |      1,398       30.24       61.97
4=Age 75-84 |      1,406       30.41       92.39
  5=Age 85+ |        352        7.61      100.00
------------+-----------------------------------
      Total |      4,623      100.00

         Year of |
      diagnosis, |
     categorised |      Freq.     Percent        Cum.
-----------------+-----------------------------------
1=Year 1990-1994 |        531       11.49       11.49
2=Year 1995-1999 |      1,009       21.83       33.31
3=Year 2000-2004 |      1,183       25.59       58.90
4=Year 2005-2009 |      1,327       28.70       87.61
```

```
5=Year 2010-2013 |        573       12.39      100.00
-----------------+-----------------------------------
           Total |      4,623      100.00
```

```
   Clinical |
   stage at |
  diagnosis |      Freq.     Percent        Cum.
------------+-----------------------------------
1=Localised |      1,828       39.54       39.54
 2=regional |      1,240       26.82       66.36
  3=Distant |      1,555       33.64      100.00
------------+-----------------------------------
      Total |      4,623      100.00
```

```
         Smoking |
          status |      Freq.     Percent        Cum.
-----------------+-----------------------------------
  0=Never smoker |      1,935       41.86       41.86
1=Current smoker |      2,688       58.14      100.00
-----------------+-----------------------------------
           Total |      4,623      100.00
```

```
  Vital status |
       at exit |      Freq.     Percent        Cum.
---------------+-----------------------------------
       0=Alive |      1,326       28.68       28.68
1=Dead: cancer |      2,544       55.03       83.71
 2=Dead: other |        753       16.29      100.00
---------------+-----------------------------------
         Total |      4,623      100.00
```

You have been provided an analysis dataset in the examination folder. The dataset is called `examdata.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/2018/examdata.csv", clear
```

SAS:

```
filename afile url "http://biostat3.net/download/exams/2018/examdata.csv";
data examdata;
    infile afile delimiter="," dsd firstobs=2;
    input id sex age agecat yydx period stage smoke surv_mm status;
run;
* or download the file locally and...;
proc import datafile="examdata.csv" out=examdata replace;
run;
```

R:

```
examdata <- read.csv("http://biostat3.net/download/exams/2018/examdata.csv")
```

Part 1

Question 1

Let's first assume that the rate of cancer-specific death is constant over time since colon cancer diagnosis. In this part you will describe how the overall (time-constant) rate of cancer-specific death differs across age and smoking status.

(a) Report the average rates by smoking status, without adjusting for any covariates. Remember to also give the units the rate is measured in [2pt].

(b) Using Poisson regression, model the overall cancer-specific mortality rates with main effects for age groups and smoking status. Report rate ratios, 95% confidence intervals and p-values, and interpret the associations for age groups and smoking status. [2pt]

(c) Propose, fit and interpret a model for the overall cancer-specific mortality rates to assess whether there is an interaction between age groups and smoking status. Formally test for whether there is evidence for an interaction. [2pt]

(d) Write out the regression equation (linear predictor) for the models in (a) and (b). (Reminder: explain your notation.) [2pt]

(e) Using the model from (b), calculate the rate ratio and 95% confidence intervals comparing the rate for a man diagnosed in age category 65-74 at diagnosis who is a smoker to a man diagnosed in the same age category who is not a smoker. [2pt]

Part 2

For this part, we are interested to determine if smoking has an effect on the colon cancer-specific death among colon cancer patients, while adjusting for potential confounders, including the time-scale time since diagnosis. Only consider the first 10 years following diagnosis.

Question 2

(a) Plot and interpret the Kaplan-Meier survival curves for colon cancer death by smoking status, over time-since diagnosis. [2pt]

(b) Formally assess and interpret whether survival varies by smoking status. [1pt]

Question 3

(a) Adjusting for time since diagnosis, fit and interpret a model for cancer-specific mortality comparing smokers and non-smokers. [2pt]

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

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

(d) Test whether the hazard ratio for smoking varies by time since diagnosis. [1pt]

(e) Fit and interpret a model that allows for non-proportional hazards for smoking status. [2pt]

(f) Fit and interpret a model for cancer-specific mortality adjusting for the time-scale (time since diagnosis), age group at diagnosis, sex and calendar period of diagnosis, where smoking status is the exposure variable of interest. Assume proportional hazards for all effects. Does the effect of smoking seem to be confounded by age, sex and calendar year of diagnosis (also motivate your answer)? [2pt]

Question 4

(a) The odds ratio from a nested case-control study can be interpreted as another epidemiological parameter, which one. [1pt]

(b) Discuss in what circumstances censoring due to emigration would be informative, and what one should do to minimise the potential bias due to potential informative censoring. [2pt]

(c) Describe the assumptions underlying the stratified Cox model, and describe situations when the stratified Cox model can be used. [2pt]