BIOSTAT III: Survival Analysis for Epidemiologists: Take-home examination

Therese Andersson

7–16 February, 2022

Instructions

- **The examination is individual-based: you are not allowed to cooperate with anyone**, although you are encouraged to consult the available literature. The examiner will use Urkund in order to assess potential plagiarism.
- The examination will be made available by noon on Wednesday 16 February 2022 and **the examination is due by 17:00 on Wednesday 23 February 2022**.
- The examination will be graded and results returned to you by Wednesday 2 March 2022.
- The examination is in two parts. To pass the examination, you need to score at least 7/13 for Part 1 focused on rates and general regression modelling and 10/19 for Part 2 on survival analysis.
- Do not write answers by hand: please use Word, LaTeX, Markdown or a similar format for your examination report and submit the report **as a PDF file**.
- Motivate all answers in your examination report. Define any notation that you use for equations. The examination report should be written in English.
- 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 or otherwise reveal your identity in the document containing the answers.**

1 Description of the data

In this exam, we will use the melanoma data presented in the course. We will specifically focus on the variable stage at diagnosis as the exposure of interest. A few extra variabes have also been created that are not included in the dataset used for the computer lab. Below is a description of the variables used in this exam, and output from stset with time since diagnosis as the time-scale and death due to melanoma as the outcome.

```
. codebook agegrp sex stage d y

-------------------------------------------------------------------------------
agegrp                                                    Age in 4 categories
-------------------------------------------------------------------------------

                  Type: Numeric (byte)
                 Label: agegrp

                 Range: [0,3]                         Units: 1
         Unique values: 4                         Missing .: 0/6,144

            Tabulation: Freq.   Numeric  Label
                        1,635         0  0-44
                        1,813         1  45-59
                        1,811         2  60-74
                          885         3  75+

-------------------------------------------------------------------------------
sex                                                                         Sex
-------------------------------------------------------------------------------

                  Type: Numeric (byte)
                 Label: sex

                 Range: [1,2]                         Units: 1
         Unique values: 2                         Missing .: 0/6,144

            Tabulation: Freq.   Numeric  Label
                        2,921         1  Male
                        3,223         2  Female

-------------------------------------------------------------------------------
stage                                          Clinical stage at diagnosis
-------------------------------------------------------------------------------

                  Type: Numeric (byte)
                 Label: stage

                 Range: [1,3]                         Units: 1
         Unique values: 3                         Missing .: 0/6,144

            Tabulation: Freq.   Numeric  Label
                        5,318         1  Localised
                          350         2  Regional
                          476         3  Distant

-------------------------------------------------------------------------------
d                             Indicator for death due to melanoma, 1=yes, 0=no
-------------------------------------------------------------------------------
```

```
                  Type: Numeric (float)

                 Range: [0,1]                         Units: 1
         Unique values: 2                         Missing .: 0/6,144

            Tabulation: Freq.  Value
                        4,505  0
                        1,639  1

------------------------------------------------------------------------------
y                               Follow-up time in exact years (#days/365.24)
------------------------------------------------------------------------------

                  Type: Numeric (float)

                 Range: [.04380681,20.961559]         Units: 1.000e-09
         Unique values: 374                       Missing .: 0/6,144

                  Mean: 6.67482
             Std. dev.: 5.18155

           Percentiles:     10%       25%       50%       75%       90%
                        1.21016   2.29438   5.29241   9.96057    14.626


. stset y, fail(d==1) exit(time 10)

Survival-time data settings

         Failure event: d==1
Observed time interval: (0, y]
     Exit on or before: time 10

------------------------------------------------------------------------------
      6,144  total observations
          0  exclusions
------------------------------------------------------------------------------
      6,144  observations remaining, representing
      1,579  failures in single-record/single-failure data
 34,501.826  total analysis time at risk and under observation
                                                At risk from t =         0
                                     Earliest observed entry t =         0
                                          Last observed exit t =        10
```

Part 1

Q 1

Below is the output from a Poisson model with death due to melanoma as the outcome and stage at diagnosis, age group at diagnosis and sex as explanatory variables.

```
. poisson d i.stage i.agegrp i.sex, exp(y)

Iteration 0:   log likelihood = -4937.8056
Iteration 1:   log likelihood = -4873.8427
Iteration 2:   log likelihood = -4873.8115
Iteration 3:   log likelihood = -4873.8115

Poisson regression                                      Number of obs =   6,144
                                                        LR chi2(6)    = 1954.56
                                                        Prob > chi2   =  0.0000
Log likelihood = -4873.8115                             Pseudo R2     =  0.1670

------------------------------------------------------------------------------
           d | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       stage |
   Regional  |   1.624063   .0751158    21.62   0.000     1.476838    1.771287
    Distant  |   2.714714   .0597173    45.46   0.000     2.597671    2.831758
             |
      agegrp |
      45-59  |   .3167896   .0713925     4.44   0.000     .1768628    .4567163
      60-74  |   .6203415   .0692078     8.96   0.000     .4846967    .7559864
        75+  |   1.124549   .0812141    13.85   0.000     .9653721    1.283725
             |
         sex |
     Female  |  -.3913987   .0510349    -7.67   0.000    -.4914253   -.2913722
       _cons |  -3.839195   .0616421   -62.28   0.000    -3.960012   -3.718379
      ln(y)  |          1  (exposure)
------------------------------------------------------------------------------

. est store A
```

a) Interpret the parameter for stage 'Distant' in the output above, including a statement about statistical significance. (2 pt)

b) What is the hazard ratio comparing a male patient with stage 'Regional' and diagnosed aged 45-59 to a male patient with stage 'Localised' and diagnosed aged 45-59? (2 pt)

c) Write out the model formulation (linear predictor) for the model above, make sure to explain your notation. (1 pt)

d) Based on the output given so far, is it possible to judge if age is a counfounder? If yes, is age a confounder (motivate your answer)? If no, why is it not possible to judge if age is a confounder based on the output above? (2 pt)

Q 2

A second Poisson model is fitted, including interaction terms between stage and age group. The model is also compared with the model fitted in Q1 using a likelihood-ratio test.

```
. poisson d i.stage##i.agegrp i.sex, exp(y)

Iteration 0:   log likelihood = -4929.6279
Iteration 1:   log likelihood = -4864.1635
Iteration 2:   log likelihood = -4864.1134
Iteration 3:   log likelihood = -4864.1134

Poisson regression                                      Number of obs =   6,144
                                                        LR chi2(12)   = 1973.96
                                                        Prob > chi2   =  0.0000
Log likelihood = -4864.1134                             Pseudo R2     =  0.1687

------------------------------------------------------------------------------
           d | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       stage |
   Regional  |   1.559185   .1645764     9.47   0.000     1.236622    1.881749
    Distant  |   2.971941   .1328707    22.37   0.000     2.711519    3.232362
             |
      agegrp |
      45-59  |   .2913583   .0903633     3.22   0.001     .1142495     .468467
      60-74  |   .6767628   .0872528     7.76   0.000     .5057504    .8477752
        75+  |   1.319568    .100775    13.09   0.000     1.122052    1.517083
             |
stage#agegrp |
   Regional #|
      45-59  |   .3387383   .2129007     1.59   0.112    -.0785395     .756016
   Regional #|
      60-74  |   -.003056   .2089844    -0.01   0.988    -.4126579    .4065459
   Regional #|
        75+  |  -.1707681   .2508509    -0.68   0.496    -.6624269    .3208907
    Distant #|
      45-59  |  -.1167497   .1716509    -0.68   0.496    -.4531793      .21968
    Distant #|
      60-74  |  -.2697977   .1660084    -1.63   0.104    -.5951681    .0555727
Distant#75+  |  -.6893975   .1930808    -3.57   0.000    -1.067829   -.3109661
             |
         sex |
      Female |  -.3836771   .0513641    -7.47   0.000    -.4843488   -.2830053
       _cons |  -3.886467   .0727479   -53.42   0.000     -4.02905   -3.743883
       ln(y) |          1  (exposure)
------------------------------------------------------------------------------

. lrtest A

Likelihood-ratio test
Assumption: A nested within .
 LR chi2(6) =  19.40
Prob > chi2 = 0.0035
```

a) Interpret the parameter for stage 'Distant' in the output above, including a statement about statistical significance. (2 pt)

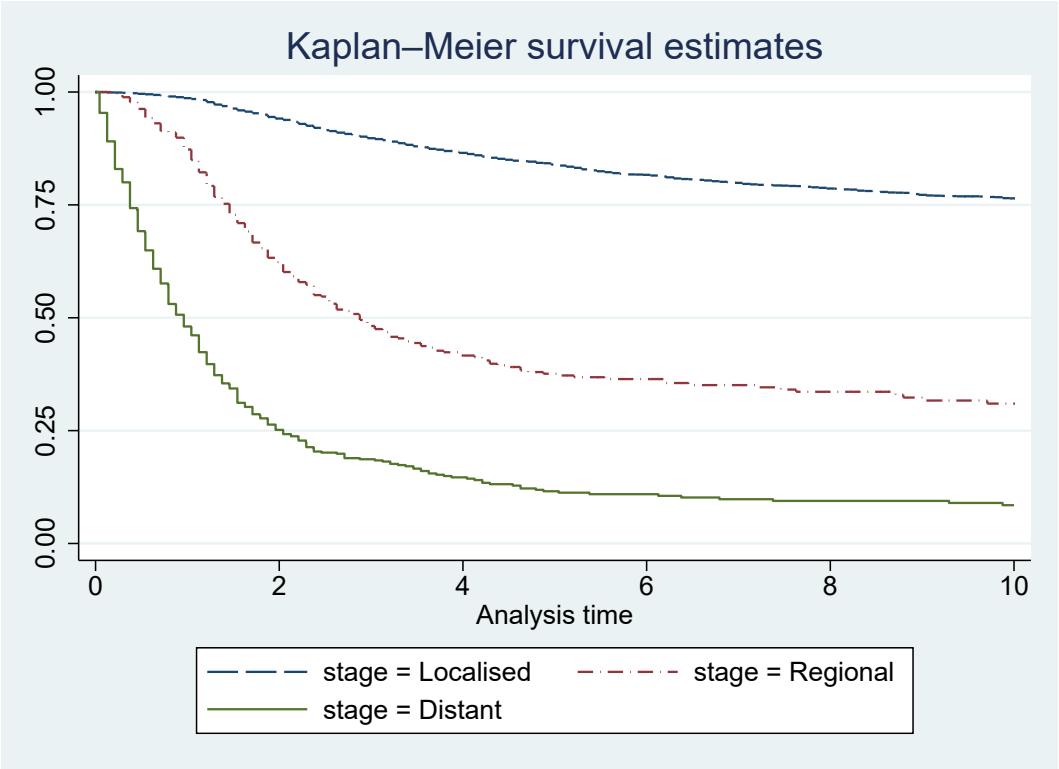
b) What is the hazard ratio comparing a male patient with stage 'Regional' and diagnosed aged 45-59 to a male patient with stage 'Localised' and diagnosed aged 45-59? (2 pt)

c) Based on the output given so far, is it possible to judge if there is effect modification by age? If yes, is there effect modification by age (motivate your answer)? If no, why is it not possible to judge if there is effect modification by age based on the output given? (2 pt)

Part 2

Q 3

Here is a Kaplan-Meier graph of the survivor function for the 3 stages, and the output from a log rank test.




```
. sts test stage

         Failure _d: d==1
   Analysis time _t: y
  Exit on or before: time 10

Equality of survivor functions
Log-rank test

           |  Observed       Expected
stage      |    events         events
-----------+-------------------------
Localised  |       960        1467.45
 Regional  |       213          66.33
  Distant  |       406          45.22
-----------+-------------------------
    Total  |      1579        1579.00

                   chi2(2) = 3443.30
                   Pr>chi2 =  0.0000
```

a) Based on the Kaplan-Meier graph, what is the 2-year survival for each of the 3 stages (approximately)? (2 pt)

b) Based on the Kaplan-Meier graph, what can you conclude about the hazard rate of death due to melanoma over time since diagnosis for the 3 stages? (3 pt)

c) Based on the log-rank test, would you conclude that there is evidence of a difference in the cancer-specific mortality across stage? (1 pt)

d) Why is it better to answer the question above using a regression model instead of a log-rank test? (2 pt)

Q 4

Below is the output from a Cox model, and test of the proportional hazards assumption based on the Schoenfelds residuals from this model.

```
. stcox i.stage i.agegrp i.sex

        Failure _d: d==1
  Analysis time _t: y
  Exit on or before: time 10

Iteration 0:   log likelihood = -13255.772
Iteration 1:   log likelihood = -12847.163
Iteration 2:   log likelihood = -12441.542
Iteration 3:   log likelihood = -12425.274
Iteration 4:   log likelihood = -12425.085
Iteration 5:   log likelihood = -12425.085
Refining estimates:
Iteration 0:   log likelihood = -12425.085

Cox regression with Breslow method for ties

No. of subjects =       6,144                           Number of obs =   6,144
No. of failures =       1,579
Time at risk    = 34,501.8262
                                                        LR chi2(6)    = 1661.37
Log likelihood = -12425.085                             Prob > chi2   =  0.0000

------------------------------------------------------------------------------
          _t | Haz. ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       stage |
   Regional  |   4.804584   .3672603    20.53   0.000     4.136093    5.581119
    Distant  |   13.76562   .8547612    42.23   0.000     12.18825    15.54713
             |
      agegrp |
      45-59  |   1.292545   .0949067     3.49   0.000     1.119296    1.492609
      60-74  |    1.63115   .1163425     6.86   0.000     1.418344    1.875885
        75+  |    2.39279   .1989125    10.50   0.000     2.033032    2.816209
             |
         sex |
     Female  |   .7050403   .0368063    -6.69   0.000     .6364691    .7809991
------------------------------------------------------------------------------

. estat phtest, detail

Test of proportional-hazards assumption

Time function: Analysis time
-----------------------------------------------------
```

	rho	chi2	df	Prob>chi2
1b.stage	.	.	1	.
2.stage	-0.12321	23.52	1	0.0000
3.stage	-0.25235	87.42	1	0.0000
0b.agegrp	.	.	1	.
1.agegrp	0.00148	0.00	1	0.9529
2.agegrp	-0.00537	0.05	1	0.8309
3.agegrp	-0.01403	0.31	1	0.5769
1b.sex	.	.	1	.
2.sex	-0.01923	0.60	1	0.4391
Global test		96.17	6	0.0000

a) Is this model equivalent to the Poisson model in question 1 (Q1)? Motivate your answer. If not, how could they the Poisson model be made more similar to the Cox model? (2 pt)

b) What is the hazard ratio comparing Regional stage to Localised stage for patients aged 75+ at diagnosis? (2 pt)

c) Write out the model formulation (linear predictor) of the model. (2 pt)

d) Is there evidence of non-proportional hazards for the covariate of interest, stage? (1 pt)

Q 5

a) Descibe a study where you would choose attained age as the time-scale. Motivate your answer. (2pt)

b) Describe two approaches for allowing for non-proportional hazards. (2 pt)