BIOSTAT III: Survival Analysis for Epidemiologists: Answers to take-home examination

Therese Andersson

8–17 February, 2021

Instructions

- 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 11/21 for Part 2 on survival analysis.

1 Description of the data

In this exam, we will use the colon cancer data presented in the course. We will specifically focus on the variable subsite as the exposure of interest (this variable has not been given a lot of focus during the course). It gives information about in which part of the colon the tumour was detected and has 4 levels, 'Coecum and ascending', 'Transverse', 'Descending and sigmoid', and 'Other and not otherwise specified (NOS)'. 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 colon cancer as the outcome.

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

                  type:  numeric (byte)
                 label:  agegrp

                 range:  [0,3]                        units:  1
         unique values:  4                        missing .:  0/13,208

            tabulation:  Freq.   Numeric  Label
                           652         0  0-44
                         2,106         1  45-59
                         5,735         2  60-74
                         4,715         3  75+

-----------------------------------------------------------------------------------------------
year8594                                                 Indicator for diagnosed during 1985-94
-----------------------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  year8594

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  0/13,208
```

```
            tabulation:  Freq.   Numeric  Label
                         5,434         0  Diagnosed 75-84
                         7,774         1  Diagnosed 85-94

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

                  type:  numeric (byte)
                 label:  sex

                 range:  [1,2]                        units:  1
         unique values:  2                        missing .:  0/13,208

            tabulation:  Freq.   Numeric  Label
                         5,455         1  Male
                         7,753         2  Female

-----------------------------------------------------------------------------------------------
subsite                                                            Anatomical subsite of tumour
-----------------------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  colonsub

                 range:  [1,4]                        units:  1
         unique values:  4                        missing .:  0/13,208

            tabulation:  Freq.   Numeric  Label
                         4,820         1  Coecum and ascending
                         2,365         2  Transverse
                         5,391         3  Descending and sigmoid
                           632         4  Other and NOS

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

                  type:  numeric (byte)
                 label:  stage

                 range:  [1,3]                        units:  1
         unique values:  3                        missing .:  0/13,208

            tabulation:  Freq.   Numeric  Label
                         6,274         1  Localised
                         1,787         2  Regional
                         5,147         3  Distant

-----------------------------------------------------------------------------------------------
d                                           Indicator for death due to colon cancer, 1=yes, 0=no
-----------------------------------------------------------------------------------------------

                  type:  numeric (float)

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  0/13,208
```

```
            tabulation:  Freq.  Value
                         6,022  0
                         7,186  1

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

                  type:  numeric (float)

                 range:  [.04380681,20.961559]        units:  1.000e-09
         unique values:  439                      missing .:  0/13,208

                  mean:   3.76028
              std. dev:    4.4187

           percentiles:        10%       25%       50%       75%       90%
                           .125945   .542109   1.87548   5.45942   10.5438

-----------------------------------------------------------------------------------------------

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

     failure event:  d == 1
obs. time interval:  (0, y]
 exit on or before:  time 10

------------------------------------------------------------------------------
     13,208  total observations
          0  exclusions
------------------------------------------------------------------------------
     13,208  observations remaining, representing
      7,122  failures in single-record/single-failure data
 43,966.874  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 colon cancer death as the outcome and subsite and age group at diagnosis as explanatory variables.

```
. poisson d i.subsite i.agegrp, exp(y)

Iteration 0:   log likelihood = -23913.572
Iteration 1:   log likelihood = -23913.443
Iteration 2:   log likelihood = -23913.443

Poisson regression                              Number of obs     =     13,208
                                                LR chi2(6)        =     759.97
                                                Prob > chi2       =     0.0000
Log likelihood = -23913.443                     Pseudo R2         =     0.0156

------------------------------------------------------------------------------------------
                     d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------------+------------------------------------------------------------------
               subsite |
            Transverse  |   .2477318   .0333812     7.42   0.000     .1823057    .3131578
Descending and sigmoid  |   .0171663   .0272406     0.63   0.529    -.0362244    .0705569
         Other and NOS  |   .1345189   .0572765     2.35   0.019      .022259    .2467788
                       |
                agegrp |
                 45-59  |   .1326942   .0638993     2.08   0.038     .0074539    .2579345
                 60-74  |   .4641152   .0586528     7.91   0.000     .3491579    .5790725
                   75+  |   .9398427   .0589442    15.94   0.000     .8243141    1.055371
                       |
                 _cons | -2.518287   .0577151   -43.63   0.000    -2.631406   -2.405167
                 ln(y) |          1  (exposure)
------------------------------------------------------------------------------------------

. est store A
```

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

   This is the log rate ratio for patients having a tumor diagnosed in 'Tranverse' region compared to patients with tumor diagnosed in 'Coecum and ascending' region, after adjusting for agegroup. This difference is statistically significant. So, patients having a tumor diagnosed in 'Transverse' region has a 28.1% (the rate ratio is exp(0.2477)=1.281) higher mortality rate than patients with tumor diagnosed in 'Coecum and ascending' region, after adjusting for agegroup.

b) Interpret the parameter for subsite 'Descending and sigmoid' in the output above, including a statement about statistical significance. (2 pt)

   This is the log rate ratio for patients having a tumor diagnosed in 'Descending and sigmoid' region compared to patients with tumor diagnosed in 'Coecum and ascending' region, after adjusting for agegroup. This difference is not statistically significant. So, patients having a tumor diagnosed in 'Descending and sigmoid' region has a 1.7% (the rate ratio is exp(0.017166)=1.017) higher mortality rate than patients with tumor diagnosed in 'Coecum and ascending' region, after adjusting for agegroup, however this is not statistically signifuicant.

c) What is the hazard ratio comparing a patient with subsite 'Transverse' and diagnosed aged 45-59 to a patient with subsite 'Coecum and ascending' and diagnosed in the youngest age group? (2 pt)

Rate for patients with 'Transverse' and aged 45-59 at diagnosis: $\lambda = \exp(\beta_0 + \beta_1 + \beta_4)$
Rate for patients with 'Coecum and ascending' and aged <45 at diagnosis: $\lambda = \exp(\beta_0)$
HR=$\exp(\beta_0 + \beta_1 + \beta_4)/\exp(\beta_0) = \exp(\beta_1) \times \exp(\beta_4) = \exp(0.2477) \times \exp(0.1327) = 1.46$

d) In this example, subsite is the exposure. We know that the distribution of age differs across subsites, and it is also known that colon cancer-specific mortality differs by age. Will this be a problem when you interpret the effect of subsite in the output above? Motivate your answer. (2 pt)

Age is a confounder in this setting. However, the model is adjusting for age, so shouldn't be a big problem in the given model, except for possible residual confounding.

Q 2

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

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

Iteration 0:   log likelihood = -23889.634
Iteration 1:   log likelihood = -23889.332
Iteration 2:   log likelihood = -23889.332

Poisson regression                              Number of obs     =     13,208
                                                LR chi2(15)       =     808.19
                                                Prob > chi2       =     0.0000
Log likelihood = -23889.332                     Pseudo R2         =     0.0166

----------------------------------------------------------------------------------------------
                           d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------------------+----------------------------------------------------------------
                     subsite |
                  Transverse |   .5488913   .1544657     3.55   0.000     .2461441    .8516385
        Descending and sigmoid |   .6811782   .1331186     5.12   0.000     .4202706    .9420859
               Other and NOS |  -.0742398   .2957184    -0.25   0.802    -.6538373    .5053576
                             |
                      agegrp |
                       45-59 |   .5470006   .1164724     4.70   0.000     .3187188    .7752824
                       60-74 |   .7903322   .1070393     7.38   0.000     .5805391    1.000125
                         75+ |   1.216963   .1070192    11.37   0.000     1.007209    1.426716
                             |
              subsite#agegrp |
            Transverse#45-59 |  -.5168524   .1779356    -2.90   0.004    -.8655997    -.168105
            Transverse#60-74 |  -.3242193   .1629553    -1.99   0.047    -.6436058   -.0048329
              Transverse#75+ |  -.2336096   .1632869    -1.43   0.153     -.553646    .0864268
Descending and sigmoid#45-59 |  -.7688978   .1514086    -5.08   0.000    -1.065653   -.4721424
Descending and sigmoid#60-74 |  -.7094681   .1393483    -5.09   0.000    -.9825857   -.4363505
  Descending and sigmoid#75+ |  -.6571303   .1402717    -4.68   0.000    -.9320578   -.3822029
         Other and NOS#45-59 |  -.2616855   .3438992    -0.76   0.447    -.9357156    .4123446
         Other and NOS#60-74 |   .1961377   .3089865     0.63   0.526    -.4094648    .8017402
           Other and NOS#75+ |   .3896897    .307874     1.27   0.206    -.2137323    .9931117
```

```
                                         |
                                   _cons |  -2.820275   .1025978   -27.49   0.000    -3.021363   -2.619187
                                   ln(y) |          1  (exposure)
------------------------------------------------------------------------------------------------------

. lrtest A

Likelihood-ratio test                                      LR chi2(9)  =     48.22
(Assumption: A nested in .)                                Prob > chi2 =    0.0000
```

a) What is the hazard ratio when comparing subsite 'Transverse' to 'Coecum and ascending' among patients diagnosed in the youngest age group. (2 pt)

$\exp(.5488913) = 1.73$

b) What is the hazard ratio when comparing subsite 'Transverse' to 'Coecum and ascending' among patients diagnosed in the ages 60-74? (2 pt)

$\exp(.5488913) * \exp(-.3242193) = 1.25$

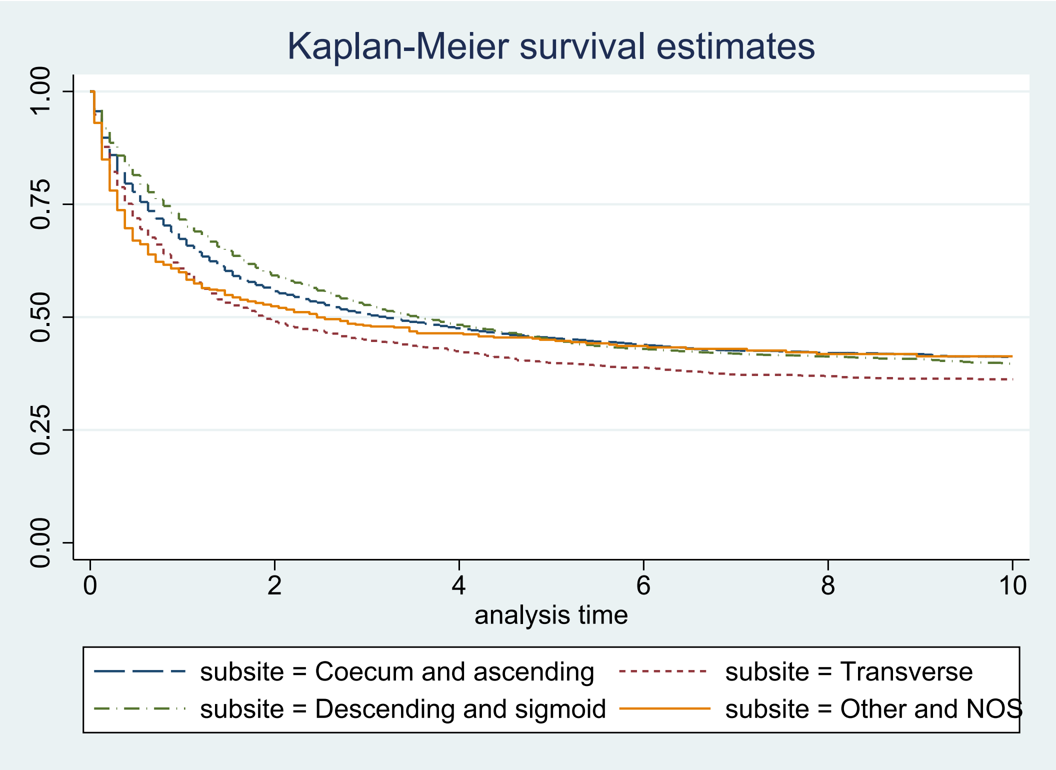
c) Is there evidence of effect modification by age? Motivate your answer. (1 pt)

Yes, the likelihood ratio test comparing the two models show a statsitically significant difference (p-value$<$0.05).

Part 2

Q 3

Here is a Kaplan-Meier graph of the survivor function for the 4 subsites, and the output from a log rank test.



```
. sts test subsite

         failure _d:  d == 1
   analysis time _t:  y
  exit on or before:  time 10


Log-rank test for equality of survivor functions

                       |   Events         Events
subsite                |  observed       expected
-----------------------+-------------------------
Coecum and ascending   |      2557        2605.52
Transverse             |      1374        1180.68
Descending and sigmoid |      2850        3021.69
Other and NOS          |       341         314.12
-----------------------+-------------------------
Total                  |      7122        7122.00

                            chi2(3) =      45.86
                            Pr>chi2 =     0.0000
```

a) Based on the Kaplan-Meier graph, what is the 1-year survival for each of the 4 subsites (approximately)? (2 pt)

Coecum and ascending: 0.67

Transverse: 0.61

Descending and sigmoid: 0.72

Other and NOS: 0.60

b) Based on the Kaplan-Meier graph, what can you conclude about the hazard rate of death due to colon cancer for the 4 subsites? (3 pt)

The hazard rate is highest within the first 2 years after diagnosis, for all subsites, and then decreases. After approximately 4-5 years the hazard rate is similar across subsites, and after 7 years the hazard rate is very low for all subsites. Within the first 2 years, 'Other and NOS' has the highest rate, and the rate for this group decreases more quickly than for the other groups. The group 'Transverse' has a higher rate than 'Coecum and ascending' and 'Descending and sigmoid' within the first 2 years.

c) Would you say that the proportional hazards assumption is reasonable? Motivate your answer. (2 pt)

No, probably not. The survival functions cross, and the effect of subsite seem to be stronger in the first years after diagnosis, since the rates are similar after 4-5 years.

d) Would you conclude that there is evidence of a difference in the cancer-specific mortality across subsites? (1 pt)

Yes, the log-rank test shows a signifcant difference between subsites (p-value < 0.05).

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

The regression model gives us an effect measure (the HR) as well as a p-value, and it allows us to adjust for confounders and allow for effect modification.

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.subsite i.agegrp

         failure _d:  d == 1
   analysis time _t:  y
  exit on or before:  time 10

Iteration 0:   log likelihood = -64476.566
Iteration 1:   log likelihood =  -64358.24
Iteration 2:   log likelihood = -64357.746
Iteration 3:   log likelihood = -64357.746
Refining estimates:
Iteration 0:   log likelihood = -64357.746

Cox regression -- Breslow method for ties

No. of subjects =       13,208                  Number of obs    =      13,208
No. of failures =        7,122
Time at risk    =  43966.87383
                                                LR chi2(6)       =      237.64
```

```
Log likelihood  =  -64357.746                   Prob > chi2     =     0.0000

------------------------------------------------------------------------------
                     _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------------------+-----------------------------------------------------
                subsite |
             Transverse |   1.213999   .0406689     5.79   0.000      1.13685    1.296384
Descending and sigmoid  |    .986004   .0269601    -0.52   0.606     .9345541    1.040286
          Other and NOS |   1.121014   .0646428     1.98   0.048     1.001213    1.255148
                        |
                 agegrp |
                 45-59  |   1.046113   .0671159     0.70   0.482     .9225032    1.186287
                 60-74  |   1.240406   .0731012     3.66   0.000     1.105095    1.392285
                   75+  |    1.60553   .0951475     7.99   0.000     1.429468    1.803278
------------------------------------------------------------------------------

. estat phtest, detail

      Test of proportional-hazards assumption

      Time:  Time
      ----------------------------------------------------------------
                  |       rho            chi2       df       Prob>chi2
      ------------+---------------------------------------------------
      1b.subsite  |            .            .        1             .
      2.subsite   |     -0.02292         3.74        1         0.0530
      3.subsite   |      0.06947        34.54        1         0.0000
      4.subsite   |     -0.04271        12.99        1         0.0003
      0b.agegrp   |            .            .        1             .
      1.agegrp    |     -0.01304         1.21        1         0.2704
      2.agegrp    |     -0.00981         0.69        1         0.4071
      3.agegrp    |     -0.03354         7.99        1         0.0047
      ------------+---------------------------------------------------
      global test |                    109.14        6         0.0000
      ----------------------------------------------------------------
```

a) Is this model equivalent to the Poisson model in question 1 (Q1)? Motivate your answer. (2 pt)

No, this model also adjusts for the time scale. The time scale is not included in the Poisson model in Q1.

b) What is the hazard ratio comparing subsite 'Other and NOS' to 'Coecum and ascending' for patients aged 75+ at diagnosis? (2 pt)

Since there is no interation between subsite and age group, the HR comparing subsite 'Other and NOS' to 'Coecum and ascending' is the same within all age groups, 1.121014.

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

$\ln(\lambda(t|X)) = \ln(\lambda_0(t)) + \beta_1 * [\text{Transverse}] + \beta_2 * [\text{Descending and sigmoid}] + \beta_3 * [\text{Other and NOS}] + \beta_4 * [\text{age 45-59}] + \beta_5 * [\text{age 60-74}] + \beta_6 * [\text{age 75+}]$

$\beta_1 = \ln(1.213)$ $\beta_2 = \ln(0.986)$ $\beta_3 = \ln(1.121)$ $\beta_4 = \ln(1.046)$ $\beta_5 = \ln(1.240)$ $\beta_6 = \ln(1.605)$

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

Yes, as the Schoenfelds residuals test rejected the hypothesis of zero slope.

e) Why would a stratified Cox model, stratifying by subsite, not be suitable in this study? (1 pt)

Because subsite is the covariate of interest whereas the stratified Cox model is suitable for data where proportional hazards assumption is violated for a factor that is not of the primary interest.

Q 5

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

For a study where it is of interest to study how the rate changes over attained age, attained age should be used as a time-scale. Otherwise, the time-scale which is suspected to have the strongest confounding effect should be chosen, so if both the exposure distribution and the rate of the event of interest differs along attained age, that should be chosen as the time-scale.

b) Describe an approach (other than stratified Cox model) of allowing for non-proportional hazards. (1 pt)

Include interaction between the covariate and the tim-scale, i.e. effect modification by time.