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

Therese Andersson

5–14 February, 2024

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 Ouriginal (`https://education.ki.se/disciplinary-matters`) in order to assess potential plagiarism.
- The examination will be made available by noon on Wednesday 14 February 2024 and **the examination is due by 17:00 on Wednesday 21 February 2024**.
- The examination is in two parts. To pass the examination, you need to score at least 9/17 for Part 1 focused on rates and general regression modelling and 11/23 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 (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 use data on breast cancer patients. The exposure variable of interest is hormonal therapy and we are interested on its effect on cancer-specific mortality (i.e. only deaths due to breast cancer are considered events, and the follow-up time for individuals that die due to other causes is censored at their time of death). Start of follow-up is at date of surgery, and the time-scale of interest is time since surgery. Follow-up is restricted to 10 years after surgery, so everyone still at risk after 10 years is censored at that point. We also have information on age at surgery and the number of positive lymph nodes (i.e. metastases in lymph nodes). Below is a description of the variables used in this exam:

```
. codebook hormon agegrp enodes d risktime

-------------------------------------------------------------------------------------------
hormon                                                   Hormonal therapy
-------------------------------------------------------------------------------------------

                Type: Numeric (byte)
               Label: adjhormo

               Range: [0,1]                        Units: 1
       Unique values: 2                        Missing .: 0/2,982

          Tabulation: Freq.   Numeric  Label
                      2,643         0  no
                        339         1  yes

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

                Type: Numeric (float)
               Label: agelabel

               Range: [0,70]                       Units: 1
       Unique values: 4                        Missing .: 0/2,982

          Tabulation: Freq.   Numeric  Label
                        712         0  <45
                      1,119        45  45-59
                        690        60  60-70
                        461        70  70+

-------------------------------------------------------------------------------------------
enodes                      Number of positive nodes (transformed as exp(-0.12 * nodes))
-------------------------------------------------------------------------------------------

                Type: Numeric (float)

               Range: [.01690747,1]                Units: 1.000e-09
       Unique values: 28                       Missing .: 0/2,982

                Mean: .795889
           Std. dev.: .263865

         Percentiles:     10%       25%       50%       75%       90%
                      .339596   .618783    .88692         1         1
```

```
--------------------------------------------------------------------------------------------------
d                        Indicator for death due to breast cancer, 1=yes, 0=no (censored)
--------------------------------------------------------------------------------------------------

                  Type: Numeric (float)

                 Range: [0,1]                          Units: 1
         Unique values: 2                          Missing .: 0/2,982

            Tabulation: Freq.  Value
                        2,049  0
                          933  1

--------------------------------------------------------------------------------------------------
risktime                          Follow-up time in exact years
--------------------------------------------------------------------------------------------------

                  Type: Numeric (float)

                 Range: [.09856263,10]                 Units: 1.000e-09
         Unique values: 1,663                      Missing .: 0/2,982

                  Mean: 6.70772
             Std. dev.: 2.92504

           Percentiles:     10%       25%       50%       75%       90%
                        2.25051   4.39973   7.22382   9.73306        10



. stset risktime, failure(d==1) exit(time 10)

Survival-time data settings

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

--------------------------------------------------------------------------
      2,982  total observations
          0  exclusions
--------------------------------------------------------------------------
      2,982  observations remaining, representing
        933  failures in single-record/single-failure data
 20,002.424  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 cancer-specific death as the outcome and hormonal therapy, age group at surgery and number of positive nodes as explanatory variables.

```
. poisson d i.hormon i.agegrp enodes, exp(risktime) irr

Iteration 0:   log likelihood = -2410.1377
Iteration 1:   log likelihood = -2409.9259
Iteration 2:   log likelihood = -2409.9259

Poisson regression                                      Number of obs =  2,982
                                                        LR chi2(5)    = 419.56
                                                        Prob > chi2   = 0.0000
Log likelihood = -2409.9259                             Pseudo R2     = 0.0801

------------------------------------------------------------------------------
           d |        IRR   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      hormon |
         yes |   .9312767    .094387    -0.70   0.482     .7634975    1.135926
             |
      agegrp |
       45-59 |   .8711058   .0726519    -1.65   0.098     .7397399      1.0258
       60-70 |   .7458675   .0731375    -2.99   0.003     .6154539    .9039156
         70+ |   .8362894   .0911259    -1.64   0.101     .6754696    1.035398
             |
      enodes |   .1007712   .0107371   -21.54   0.000      .081779    .1241742
       _cons |   .3080836     .03005   -12.07   0.000     .2544741    .3729869
------------------------------------------------------------------------------

. est store A
```

a) Interpret the parameter for hormonal therapy (‘hormon’) in the output above, including a statement about statistical significance. (2 p)

b) Interpret the parameter for age group ’60-70’ in the output above, including a statement about statistical significance. (2 p)

c) Write out the model formulation (linear predictor) for the model above, make sure to explain your notation. (2 p)

d) What is the hazard ratio comparing a patient who received hormonal therapy and had surgery aged ‘60-70’ to a patient who had no hormonal therapy and had surgery aged ‘70+’? For this comparison assume that both patients had the same number of positive nodes. (2 p)

e) Based on the output given so far, is it possible to judge if age is a confounder? 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 p)

Q 2

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

```
. poisson d i.hormon##i.agegrp enodes , exp(risktime) irr

Iteration 0:   log likelihood = -2409.7726
Iteration 1:   log likelihood = -2409.5562
Iteration 2:   log likelihood = -2409.5562

Poisson regression                                     Number of obs =  2,982
                                                       LR chi2(8)    = 420.30
                                                       Prob > chi2   = 0.0000
Log likelihood = -2409.5562                            Pseudo R2     = 0.0802

------------------------------------------------------------------------------
           d |        IRR   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      hormon |
        yes  |   .7148819   .3603596    -0.67   0.506     .2661695     1.92004
             |
      agegrp |
       45-59 |   .8611858    .073812    -1.74   0.081     .7280155    1.018716
       60-70 |   .7346041   .0767013    -2.95   0.003     .5986568    .9014232
         70+ |    .850833   .0980707    -1.40   0.161     .6787832    1.066492
             |
hormon#agegrp |
   yes#45-59 |   1.370151   .7234092     0.60   0.551     .4868028    3.856416
   yes#60-70 |    1.35987    .721851     0.58   0.563     .4804626    3.848888
     yes#70+ |   1.160671   .6413504     0.27   0.787     .3929725    3.428123
             |
      enodes |    .101147   .0107904   -21.48   0.000     .0820629    .1246692
       _cons |   .3087932   .0301445   -12.04   0.000     .2550194    .3739058
------------------------------------------------------------------------------

. est store B
. lrtest A B

Likelihood-ratio test
Assumption: A nested within B

 LR chi2(3) =   0.74
Prob > chi2 = 0.8639
```

a) Interpret the parameter for hormonal therapy (‘hormon’) in the output above, including a statement about statistical significance. (2 p)

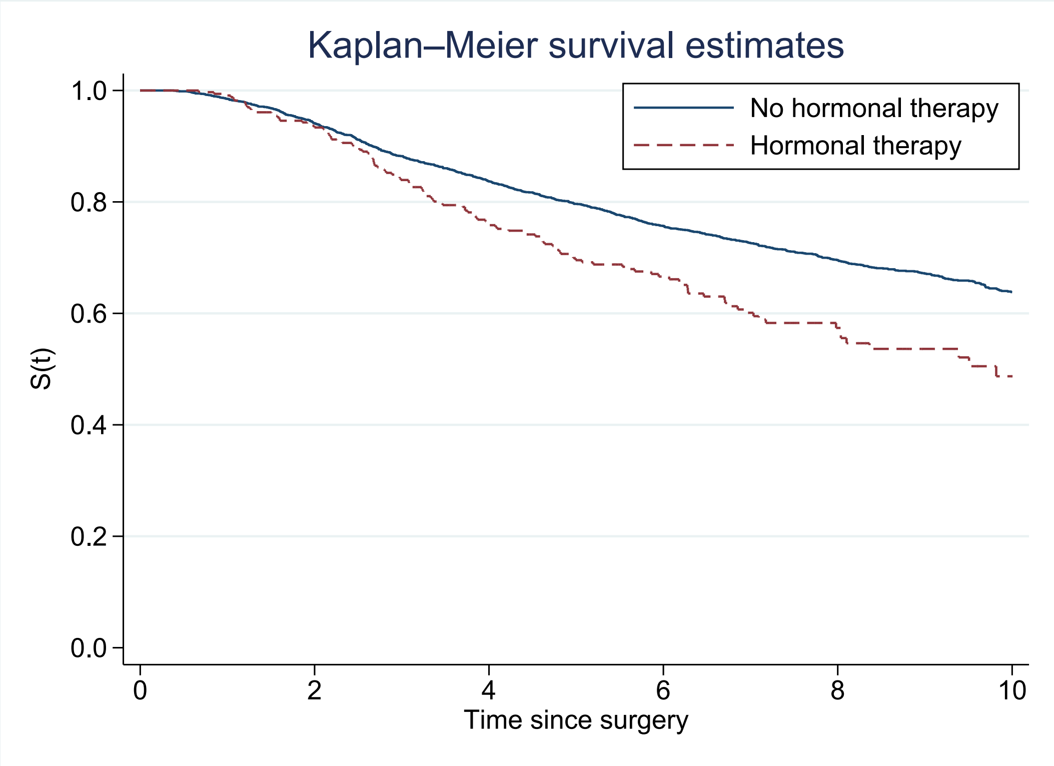
b) What is the hazard ratio comparing a patient who received hormonal therapy and had surgery aged ‘60-70’ to a patient who had no hormonal therapy and had surgery aged ’60-70’? For this comparison assume that both patients had the same number of positive nodes. (2 p)

c) Is there evidence of effect modification by age on the effect of hormonal therapy? Motivate your answer. (3 p)

Part 2

Q 3

Below is a Kaplan-Meier graph of the survivor function for the 2 treatment groups, and the output from a log rank test.




```
. sts test hormon

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

Equality of survivor functions
Log-rank test

       |  Observed       Expected
hormon |    events         events
-------+-------------------------
    no |       808         847.10
   yes |       125          85.90
-------+-------------------------
 Total |       933         933.00

                 chi2(1) =  19.72
                 Pr>chi2 = 0.0000
```

a) Based on the Kaplan-Meier graph, what is the 5-year survival for each of the 2 treatment groups (approximately)? (2 p)

b) Based on the Kaplan-Meier graph, which of the 2 treatment groups has a better survival? (2 p)

c) Based on the Kaplan-Meier graph, what can you conclude about the hazard rate of death for each treatment group? To get full marks on this question the answer has to include both the general pattern of the shape of the hazard functions as well as information on differences between the groups. (4 p)

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

e) Based on the log-rank test, is there evidence of a difference in cancer-specific mortality between hormonal therapy and no hormonal therapy? (1 p)

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.hormon i.agegrp enodes

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

Iteration 0:   log likelihood =  -7142.741
Iteration 1:   log likelihood = -6973.6741
Iteration 2:   log likelihood = -6920.5774
Iteration 3:   log likelihood = -6920.4749
Iteration 4:   log likelihood = -6920.4749
Refining estimates:
Iteration 0:   log likelihood = -6920.4749

Cox regression with Breslow method for ties

No. of subjects =      2,982                        Number of obs =  2,982
No. of failures =        933
Time at risk    = 20,002.4244
                                                    LR chi2(5)    = 444.53
Log likelihood = -6920.4749                         Prob > chi2   = 0.0000

------------------------------------------------------------------------------
          _t | Haz. ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      hormon |
        yes  |   .9360382   .0950292    -0.65   0.515     .7671446    1.142115
             |
      agegrp |
      45-59  |   .8660101   .0722287    -1.72   0.085     .7354097    1.019804
      60-70  |   .7406609   .0726702    -3.06   0.002     .6110876    .8977086
        70+  |   .8488208   .0924905    -1.50   0.133     .6855927    1.050911
             |
      enodes |   .0918915    .009867   -22.23   0.000      .074452    .1134161
------------------------------------------------------------------------------
```

```
. // Schoenfeld residuals
. estat phtest, detail

Test of proportional-hazards assumption

Time function: Analysis time
--------------------------------------------------------
             |       rho     chi2       df    Prob>chi2
-------------+------------------------------------------
   0b.hormon |         .        .        1            .
    1.hormon |   0.04241     1.67        1       0.1965
   0b.agegrp |         .        .        1            .
   45.agegrp |   0.02390     0.53        1       0.4663
   60.agegrp |   0.00089     0.00        1       0.9784
   70.agegrp |   0.05015     2.35        1       0.1251
      enodes |   0.09249     6.72        1       0.0095
-------------+------------------------------------------
 Global test |              10.25        5       0.0683
--------------------------------------------------------
```

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

b) Write out the model formulation (linear predictor) of the Cox model. (2 p)

c) What is the hazard ratio comparing hormonal therapy to no hormonal therapy for patients within the same age category at surgery and the same number of positive nodes (enodes)? (2 p)

d) Is there evidence of non-proportional hazards for any of the covariates in the model? Motivate your answer. (2 p)

Q 5

a) Give two reasons why it can be better to explore differences in survival outcomes using a regression model instead of a log-rank test. (2 p)

b) Explain why informative censoring might be a problem when interpreting the Kaplan-Meier graph in Q3, but less of a problem in the Cox regression in Q4. Motivate your answer (2 p)