BIOSTAT III: Survival Analysis for Epidemiologists:

Take-home examination

Anna Johansson

3-12 February, 2025

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 12:00 on Wednesday 12 February 2025 and the examination is due by 17:00 on Wednesday 19 February 2025.
- 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 **13/25 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.

Description of the data

In this exam we use data on **breast cancer patients**. The exposure variable of interest is **tumour grade** (which is a tumour marker) and we are interested on its effect on all-cause mortality. 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, the size of the tumour and year of surgery. Below is a description of the variables used in this exam:

```
. codebook grade agegrp size yyyy d risktime

-------------------------------------------------------------------------------------------
grade                                                                  Differention grade
-------------------------------------------------------------------------------------------

                  Type: Numeric (float)
                 Label: grlab

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

            Tabulation: Freq.   Numeric  Label
                          794         0  Low grade
                        2,188         1  High grade

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

                  Type: Numeric (float)
                 Label: agelab

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

            Tabulation: Freq.   Numeric  Label
                          712         0  0-44
                        1,119        45  45-59
                          690        60  60-69
                          461        70  70+

-------------------------------------------------------------------------------------------
size                                                         Tumour size, 3 classes (t)
-------------------------------------------------------------------------------------------

                  Type: Numeric (byte)
                 Label: size

                 Range: [1,3]                         Units: 1
         Unique values: 3                         Missing .: 0/2,982

            Tabulation: Freq.   Numeric  Label
                        1,387         1  <=20 mm
                        1,291         2  >20-50mmm
                          304         3  >50 mm
```

```
-------------------------------------------------------------------------------
yyyy                    Year rescaled (year-1977), i.e. yyyy=1 means 1978 and so on
-------------------------------------------------------------------------------

                  Type: Numeric (float)

                 Range: [1,16]                        Units: 1
         Unique values: 16                         Missing .: 0/2,982

                  Mean: 11.1613
             Std. dev.: 3.03548

           Percentiles:     10%       25%       50%       75%       90%
                             7         9        11        13        15

-------------------------------------------------------------------------------
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
                        1,811  0
                        1,171  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
      1,171  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:

Question 1

Below is the output from a Poisson model with all-cause deaths as the outcome and grade, age group at surgery, year of surgery and size as explanatory variables.

```
. poisson d i.grade ib45.agegrp i.size yyyy, exp(risktime) irr

Iteration 0:  Log likelihood = -2867.1698
Iteration 1:  Log likelihood = -2867.1308
Iteration 2:  Log likelihood = -2867.1308

Poisson regression                                      Number of obs =  2,982
                                                        LR chi2(7)    = 310.13
                                                        Prob > chi2   = 0.0000
Log likelihood = -2867.1308                             Pseudo R2     = 0.0513

------------------------------------------------------------------------------
           d |        IRR   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       grade |
 High grade  |   1.499624   .1114577     5.45   0.000     1.296336     1.73479
             |
      agegrp |
       0-44  |   1.036292   .0835075     0.44   0.658     .8848907    1.213597
      60-69  |   1.081431    .086037     0.98   0.325     .9252913    1.263919
        70+  |   1.569462    .128054     5.52   0.000      1.33752    1.841625
             |
        size |
  >20-50mmm  |   1.864156    .124842     9.30   0.000     1.634848    2.125626
     >50 mm  |   3.145683   .2829492    12.74   0.000     2.637247    3.752142
             |
        yyyy |   .9703967   .0095337    -3.06   0.002     .9518898    .9892635
       _cons |   .0348141   .0046658   -25.05   0.000     .0267718    .0452724
------------------------------------------------------------------------------

. est store A
```

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

b) Interpret the parameter for age group ’60-69’ 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 with a high grade tumour and had surgery aged ‘60-69’ to a patient who had a low grade tumour and had surgery aged ‘0-44’? For this comparison assume that both patients were diagnosed in the same year and with the same tumour size. (2 p)

e) Based on the output given so far, is it possible to judge if age or year are confounders? If yes, are age or year confounders (motivate your answer)? If no, why is it not possible to judge if age or year are confounders based on the output above? (2 p)

Question 2

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

```
. poisson d i.grade##ib45.agegrp i.size yyyy , exp(risktime) irr

Iteration 0:  Log likelihood = -2861.7554
Iteration 1:  Log likelihood = -2861.7148
Iteration 2:  Log likelihood = -2861.7148

Poisson regression                                      Number of obs =  2,982
                                                        LR chi2(10)   = 320.97
                                                        Prob > chi2   = 0.0000
Log likelihood = -2861.7148                             Pseudo R2     = 0.0531

------------------------------------------------------------------------------
           d |        IRR   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       grade |
  High grade |   2.088724   .2816913     5.46   0.000     1.603561    2.720674
             |
      agegrp |
        0-44 |   1.399104   .2630725     1.79   0.074     .9678305    2.022557
       60-69 |   1.718762   .3101173     3.00   0.003     1.206793    2.447927
         70+ |   2.394764   .4308472     4.85   0.000     1.683144    3.407252
             |
 grade#agegrp |
High grade#0-44 |   .6882685   .1432058    -1.80   0.073     .4577739     1.03482
High grade#60-69 |   .5621437   .1129845    -2.87   0.004     .3791102    .8335453
High grade#70+ |   .5898193   .1184331    -2.63   0.009     .3979252    .8742517
             |
        size |
   >20-50mmm |   1.858132   .1245184     9.25   0.000     1.629428    2.118936
      >50 mm |   3.119618   .2804678    12.65   0.000      2.61562    3.720731
             |
        yyyy |   .9700884   .0095255    -3.09   0.002     .9515971    .9889389
       _cons |   .0268524   .0044012   -22.07   0.000     .0194745    .0370254
------------------------------------------------------------------------------

. est store B

. lrtest A B

Likelihood-ratio test
Assumption: A nested within B

 LR chi2(3) =  10.83
Prob > chi2 = 0.0127
```

a) Interpret the parameter for tumour grade ('grade') in the output above, including a statement about statistical significance. (2 p)

b) What is the hazard ratio comparing a patient with a high grade tumour and had surgery aged '60-69' to a patient who had a low grade tumour and had surgery aged '60-69'? For this comparison assume that both patients were diagnosed in the same year with the same tumour size. (2 p)

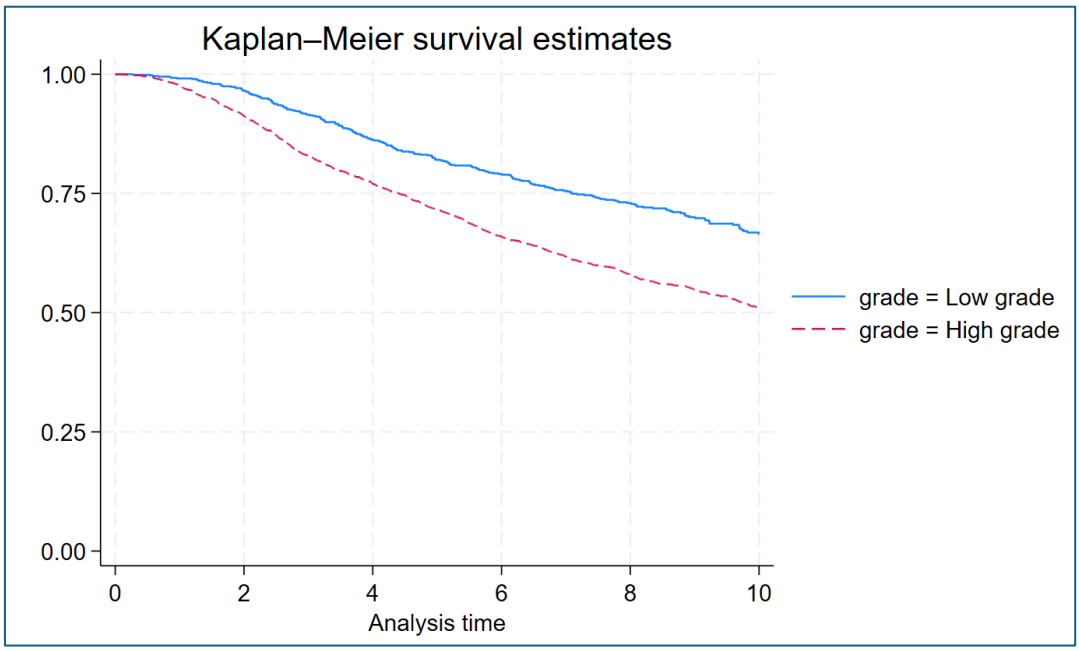
c) Is there evidence of effect modification by age on the effect of tumour grade? Motivate your answer. (1 p)

d) Interpret the parameter for year ('yyyy') in the output above, including a statement about statistical significance. (2p)

PART 2:

Question 3

Below is a Kaplan-Meier graph of the survivor function for the two groups with low and high grade tumours, and the output from a log rank test comparing the two groups.




```
. sts test grade

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

Equality of survivor functions
Log-rank test

           |  Observed       Expected
grade      |    events         events
-----------+--------------------------
 Low grade |       231         344.75
High grade |       940         826.25
-----------+--------------------------
     Total |      1171        1171.00

                   chi2(1) =  53.30
                   Pr>chi2 = 0.0000
```

a) Based on the Kaplan-Meier graph, what is the 4-year survival for the low grade group and the high grade group (approximately)? (2 p)

b) Based on the Kaplan-Meier graph, which of the low and high grade groups has the better survival? (2 p)

c) Based on the Kaplan-Meier graph, what can you conclude about the hazard rate of death for each of the low and high grade groups? (2 p)

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

e) Is there evidence of a difference in all-cause mortality rates between low and high grade? (1 p)

Question 4

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

```
. stcox i.grade ib45.agegrp i.size yyyy

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

Iteration 0:  Log likelihood = -8957.8518
Iteration 1:  Log likelihood = -8816.6343
Iteration 2:  Log likelihood = -8795.5675
Iteration 3:  Log likelihood = -8795.3365
Iteration 4:  Log likelihood = -8795.3364
Refining estimates:
Iteration 0:  Log likelihood = -8795.3364

Cox regression with Breslow method for ties

No. of subjects =       2,982                           Number of obs =  2,982
No. of failures =       1,171
Time at risk    = 20,002.4244
                                                        LR chi2(7)    = 325.03
Log likelihood = -8795.3364                             Prob > chi2   = 0.0000

------------------------------------------------------------------------------
          _t | Haz. ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       grade |
 High grade  |   1.514448   .1125549     5.58   0.000     1.309159    1.751928
             |
      agegrp |
       0-44  |   1.036643   .0835456     0.45   0.655     .8851746    1.214031
      60-69  |   1.081929    .086092     0.99   0.322     .9256914    1.264536
        70+  |   1.587091   .1296578     5.65   0.000     1.352268    1.862691
             |
        size |
  >20-50mmm  |   1.890141   .1267021     9.50   0.000     1.657431    2.155524
     >50 mm  |    3.28927   .2967074    13.20   0.000     2.756241    3.925381
             |
        yyyy |   .9714572    .009716    -2.90   0.004     .9525996    .9906881
------------------------------------------------------------------------------
```

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

Test of proportional-hazards assumption

Time function: Analysis time
------------------------------------------------------------
             |        rho     chi2       df     Prob>chi2
-------------+----------------------------------------------
    0b.grade |          .        .        1             .
     1.grade |   -0.03458     1.42        1        0.2340
    0.agegrp |   -0.02674     0.84        1        0.3591
  45b.agegrp |          .        .        1             .
   60.agegrp |   -0.00143     0.00        1        0.9608
   70.agegrp |    0.06536     5.15        1        0.0233
     1b.size |          .        .        1             .
      2.size |   -0.03542     1.49        1        0.2230
      3.size |   -0.06033     4.35        1        0.0369
        yyyy |    0.02539     0.79        1        0.3743
-------------+----------------------------------------------
 Global test |               14.46        7        0.0436
------------------------------------------------------------
```

a) Is this model equivalent to the Poisson model in Question 1? 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 high grade to low grade for patients within the same age category at surgery, the same tumour size and the same year at surgery? (2 p)

d) For which time period since surgery is this hazard ratio in c) valid (the maximum follow-up is 10 years)? Motivate your answer? (2p)

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

f) Is any of the models in Question 1, Question 2 and Question 3 more suitable than the other two? Would you fit an alternative model to these data? Motivate your answer. (2 p)

Question 5

a) In which situations would a log-rank test be equally good as a regression model to compare the survival between two or more groups? Motivate your answer. (2 p)

b) Describe a situation with multiple timescales and how you would choose the main timescale in your Cox regression model in such a situation. Motivate your answer. (2 p)