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

Mark Clements

10–19 February, 2020

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 19 February 2020 and **the examination is due by 17:00 on Wednesday 26 February 2020**.
- The examination will be graded and results returned to you by Wednesday 4 March 2020.
- The examination is in two parts. You need to score at least 8/15 for Part 1 focused on rates and general regression modelling and 13/24 in Part 2 on survival analysis to pass the examination.
- Do not write answers by hand: please use Word, LaTeX 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.**

Part 1

The **DMepi2** dataset includes simulated data on all cause mortality rates for those with and without diabetes in Denmark for 1996–2015. The dataset has the following columns:

sex a factor with levels 1=M, 2=F

A One-year age class, 0–99 years

P Calendar year, 1996–2016

diab Indicator for persons with diabetes (1=yes, 0=no)

Y Person-years

D Number of deaths

R Rates (=D/Y)

Q1

(a) The age-specific mortality rates by sex and diabetes status for 2016 are shown in Figure 1. Carefully describe the pattern of rates by age, sex and diabetes status. (2 pts)

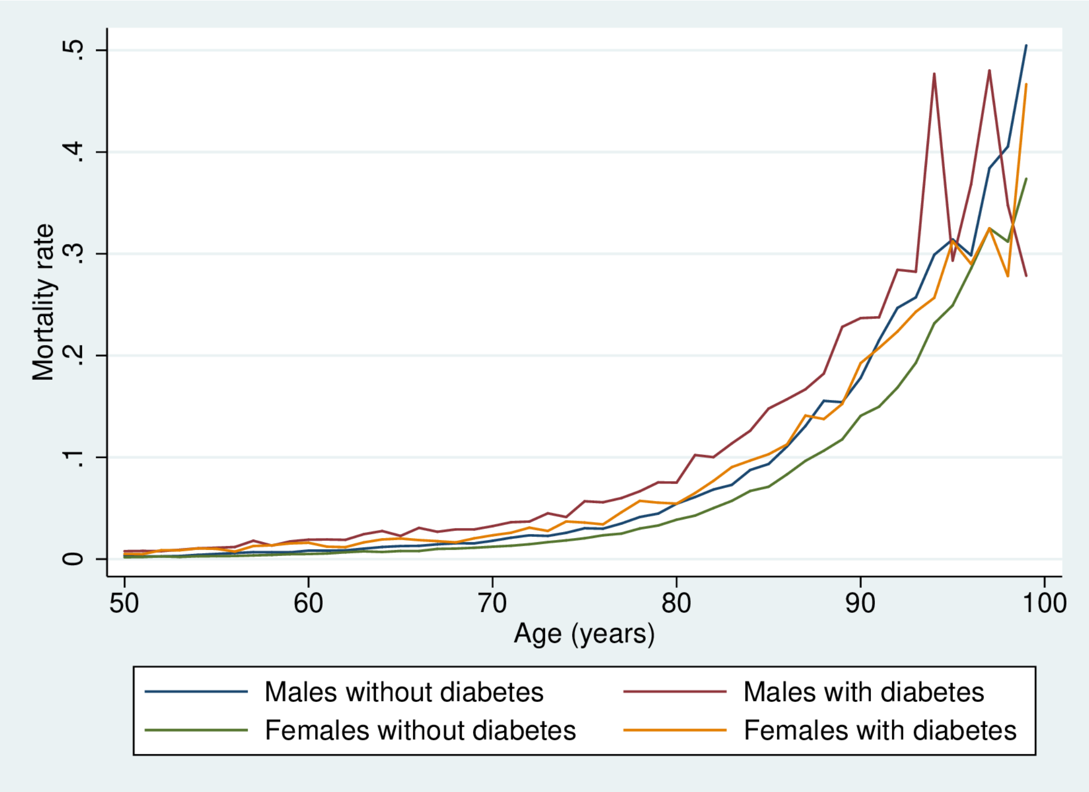

Figure 1: Age-specific mortality rates by sex and presence or absence of diabetes, Denmark 2016.

The following code and output is used to model the mortality rates by diabetes status for males and females separately for the 2016 calendar year:

```
use DMepi2, clear
keep if P==2016
poisson D A diab if sex==1, exp(Y)

(7,997 observations deleted)
Poisson regression                                Number of obs     =        199
                                                  LR chi2(2)        =   72116.18
                                                  Prob > chi2       =     0.0000
Log likelihood = -936.90877                       Pseudo R2         =     0.9747
------------------------------------------------------------------------------
           D |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           A |   .0989664   .0004939   200.36   0.000     .0979983    .0999345
        diab |   .4968066   .0150224    33.07   0.000     .4673633    .5262499
       _cons |  -10.80233   .0372321  -290.13   0.000     -10.8753   -10.72936
       ln(Y) |          1  (exposure)
------------------------------------------------------------------------------

poisson D A diab if sex==2, exp(Y)
```

```
Poisson regression                                Number of obs     =        199
                                                  LR chi2(2)        =   80738.73
                                                  Prob > chi2       =     0.0000
Log likelihood = -873.46624                       Pseudo R2         =     0.9788
------------------------------------------------------------------------------
           D |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           A |   .1076621   .0005173   208.11   0.000     .1066481     .108676
        diab |    .451282   .0164922    27.36   0.000     .4189578    .4836061
       _cons |  -11.78155   .0415506  -283.55   0.000    -11.86299   -11.70011
       ln(Y) |          1  (exposure)
------------------------------------------------------------------------------
```

(b) Write out the regression model for males. As a reminder, please explain all of your notation. (2 pts)

(c) What are the mortality rate ratios and 95% confidence intervals for those with diabetes compared with those without diabetes for (i) males and (ii) females? (2 pts)

The following interaction model and linear combination can be used to compare the mortality rate ratio of diabetes for males with the mortality rate ratio of diabetes for females. As a reminder, `baselevels` adds the base or reference level for a factor variable to the output.

```
poisson D A diab##sex, exp(Y) baselevels
lincom 1.diab + 1.diab#2.sex
```

```
Poisson regression                                Number of obs     =        398
                                                  LR chi2(4)        =  152711.67
                                                  Prob > chi2       =     0.0000
Log likelihood = -1884.4039                       Pseudo R2         =     0.9759
------------------------------------------------------------------------------
           D |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           A |    .103223   .0003568   289.26   0.000     .1025236    .1039224
             |
        diab |
          0  |          0  (base)
          1  |   .4854449   .0149829    32.40   0.000     .4560789    .5148108
             |
         sex |
          M  |          0  (base)
          F  |  -.3127235   .0100123   -31.23   0.000    -.3323473   -.2930997
             |
    diab#sex |
        1#F  |  -.0245248   .0222395    -1.10   0.270    -.0681135    .0190639
             |
       _cons |  -11.11888   .0275994  -402.87   0.000    -11.17298   -11.06479
       ln(Y) |          1  (exposure)
------------------------------------------------------------------------------
 ( 1)  [D]1.diab + [D]1.diab#2.sex = 0
------------------------------------------------------------------------------
           D |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
        (1) |   .4609201   .0164799    27.97   0.000     .4286201    .4932201
------------------------------------------------------------------------------
```

(d) Write out the regression equation for the interaction model. (1pt)

(e) What are the mortality rate ratios and 95% confidence intervals for those with diabetes compared with those without diabetes for (i) males and (ii) females? Why are these estimates different to the estimates in (c)? (2pts)

(f) Formally test for whether the two mortality rate ratios for males and females in **(e)** are different. Explain how you undertook the test and interpret the findings. (2pts)

We now model calendar period as a continuous, linear effect using a main effects model:

```
use DMepi2, clear
poisson D A sex diab P, exp(Y) baselevels

Poisson regression                              Number of obs     =      8,395
                                                LR chi2(4)        = 3398186.17
                                                Prob > chi2       =     0.0000
Log likelihood = -37872.375                     Pseudo R2         =     0.9782
------------------------------------------------------------------------------
           D |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           A |   .0978487   .0000704  1390.13   0.000     .0977108    .0979867
         sex |  -.3675351   .0018946  -194.00   0.000    -.3712484   -.3638218
        diab |   .5393626   .0026551   203.14   0.000     .5341586    .5445665
           P |  -.0253965   .0001544  -164.52   0.000     -.025699   -.0250939
       _cons |   40.87618   .3096228   132.02   0.000     40.26933    41.48303
       ln(Y) |          1  (exposure)
------------------------------------------------------------------------------
```

(g) Write out the regression equation for this model. (1pt)

(h) How would you interpret the parameter `P` and its 95% confidence interval? (1pt)

(i) How would you interpret the parameter `_cons`? Why is the value so large? (2pts)

Part 2

Q2

We now use data from the German Breast Cancer Study Group (GBCSG) on a randomised study of hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients (see `https://doi.org/10.1200/JCO.1994.12.10.2086`). The event considered was time to recurrence of breast cancer or death due to breast cancer ("recurrence-free survival"). The main study found no effect associated with the duration of chemotherapy on recurrence-free survival. The code-book for some of the dataset is shown below:

```
use brcancer, clear
keep id hormon x1 x6 rectime censrec
egen x1cat = cut(x1), at(0,45,60,80) label
egen x6cat = cut(x6), at(0,20,2380) label
codebook
```

```
(German breast cancer data)
(2 missing values generated)
(1 missing value generated)
-------------------------------------------------------------------------------
id                                                                Individual ID
-------------------------------------------------------------------------------
                  type:  numeric (int)
                 range:  [1,686]                      units:  1
         unique values:  686                      missing .:  0/686
                  mean:       343.5
              std. dev:     198.175
           percentiles:        10%       25%       50%       75%       90%
                                69       172     343.5       515       618
-------------------------------------------------------------------------------
hormon                                                         hormonal therapy
-------------------------------------------------------------------------------
                  type:  numeric (byte)
                 label:  hormon
                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  0/686
            tabulation:  Freq.   Numeric  Label
                           440         0  Standard treatment
                           246         1  Hormonal treatment
-------------------------------------------------------------------------------
x1                                                                   age, years
-------------------------------------------------------------------------------
                  type:  numeric (byte)
                 range:  [21,80]                      units:  1
         unique values:  54                       missing .:  0/686
                  mean:     53.0525
              std. dev:     10.1207
           percentiles:        10%       25%       50%       75%       90%
                                40        46        53        61        65
-------------------------------------------------------------------------------
x6                                                  progesterone receptor, fmol
-------------------------------------------------------------------------------
                  type:  numeric (int)
                 range:  [0,2380]                     units:  1
         unique values:  242                      missing .:  0/686
                  mean:     109.996
              std. dev:     202.332
           percentiles:        10%       25%       50%       75%       90%
                                 0         7      32.5       132       312
-------------------------------------------------------------------------------
rectime                                     recurrence-free survival time, days
-------------------------------------------------------------------------------
                  type:  numeric (int)
                 range:  [8,2659]                     units:  1
         unique values:  574                      missing .:  0/686
                  mean:     1124.49
              std. dev:     642.792
           percentiles:        10%       25%       50%       75%       90%
                               322       567      1084      1685      2014
-------------------------------------------------------------------------------
censrec                                                      censoring indicator
-------------------------------------------------------------------------------
                  type:  numeric (byte)
                 label:  event
                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  0/686
           tabulation:  Freq.   Numeric  Label
                           387         0  censored
                           299         1  event
-------------------------------------------------------------------------------
x1cat                                                               (unlabeled)
-------------------------------------------------------------------------------
                  type:  numeric (float)
                 label:  x1cat
                 range:  [0,2]                        units:  1
         unique values:  3                        missing .:  2/686
           tabulation:  Freq.   Numeric  Label
                           131         0  0-
                           344         1  45-
                           209         2  60-
                             2         .
-------------------------------------------------------------------------------
x6cat                                                               (unlabeled)
-------------------------------------------------------------------------------
                  type:  numeric (float)
                 label:  x6cat
                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  1/686
           tabulation:  Freq.   Numeric  Label
                           269         0  0-
                           416         1  20-
                             1         .
```

(a) For this dataset, we have the time from randomisation to recurrence or breast cancer death. Assume we also had (i) the date of birth, (ii) date of cancer diagnosis, (iii) date of randomisation and (iv) date of recurrence or death. Discuss which time scales you could use for your analysis, describing their advantages and disadvantages. (2pts)

We now **stset** for the time from randomisation to time of recurrence or death – that is, we are modelling for recurrence-free survival. There were 299 events and the event times are in days from randomisation.

(b) The Kaplan-Meier estimators for the survival functions by hormonal treatment are shown in Figure 2. Carefully describe and interpret the two survival curves. (2pts)

```
stset rectime, fail(censrec==1)
sts graph, by(hormon) title("")

     failure event:  censrec == 1
obs. time interval:  (0, rectime]
 exit on or before:  failure
------------------------------------------------------------------------------
        686  total observations
          0  exclusions
------------------------------------------------------------------------------
        686  observations remaining, representing
        299  failures in single-record/single-failure data
    771,400  total analysis time at risk and under observation
                                                at risk from t =         0
                                     earliest observed entry t =         0
                                          last observed exit t =     2,659

         failure _d:  censrec == 1
   analysis time _t:  rectime
```

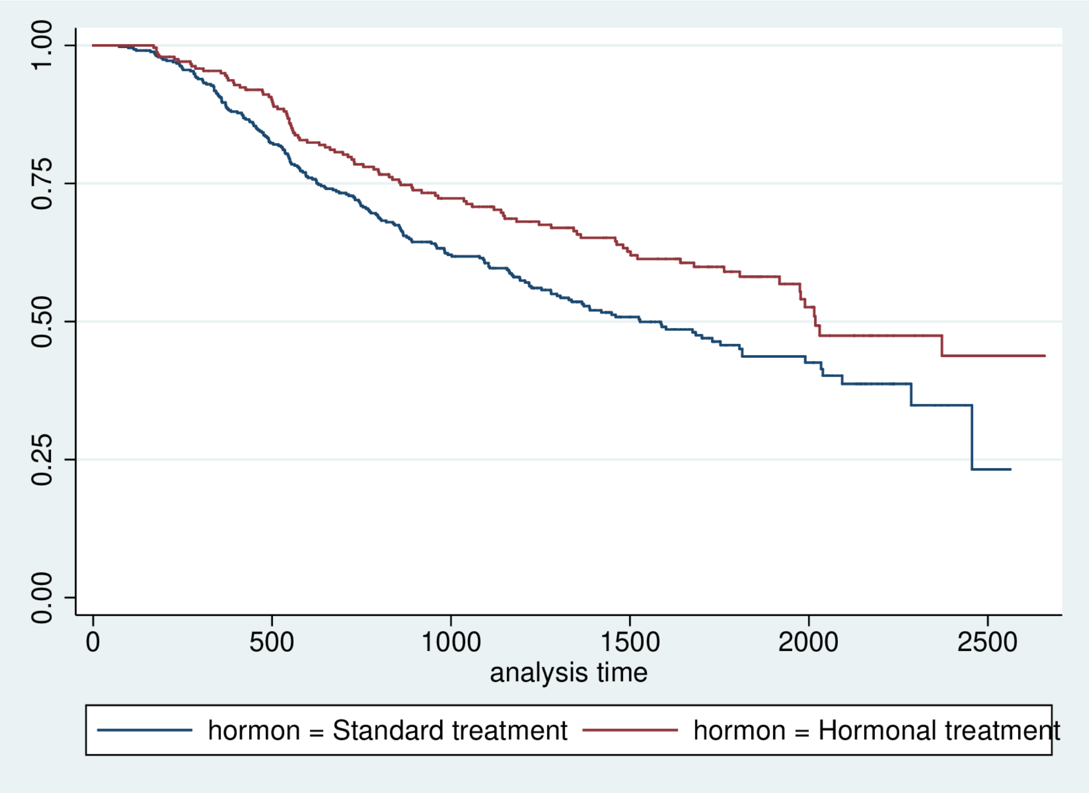

Figure 2: Kaplan-Meier survival curves by hormonal treatment, German Breast Cancer Study Group

(c) For the following log-rank test, state the null hypothesis and interpret the test. (1pt)

```
sts test hormon

         failure _d:  censrec == 1
   analysis time _t:  rectime
Log-rank test for equality of survivor functions
------------------------------------------------
                   |   Events         Events
hormon             |  observed       expected
-------------------+-------------------------
```

```
Standard treatment |       205        180.34
Hormonal treatment |        94        118.66
-------------------+-------------------------
Total              |       299        299.00
                        chi2(1) =       8.56
                        Pr>chi2 =     0.0034
```

(d.i) Write out the regression equation for the Cox model specified in the following code and output. (2pts)

(d.ii) Based on the following output, discuss whether there is any evidence that hormonal treatment is associated with recurrence-free survival. (2pts)

```
stcox i.x1cat i.x6cat hormon, baselevels nohr

         failure _d:  censrec == 1
   analysis time _t:  rectime
Refining estimates:
Cox regression -- Breslow method for ties
No. of subjects =          684                     Number of obs    =         684
No. of failures =          298
Time at risk    =       770171
                                                   LR chi2(4)       =       55.37
Log likelihood  =   -1753.6988                     Prob > chi2      =      0.0000
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       x1cat |
         0-  |          0  (base)
        45-  |  -.3036346   .1501781    -2.02   0.043    -.5979782   -.0092909
        60-  |  -.1439799   .1642114    -0.88   0.381    -.4658284    .1778685
             |
       x6cat |
         0-  |          0  (base)
        20-  |  -.7635767   .1164474    -6.56   0.000    -.9918095   -.5353439
             |
      hormon |  -.3453868   .1274155    -2.71   0.007    -.5951166   -.0956569
------------------------------------------------------------------------------
```

(e) Based on the following Schoenfeld residuals table, is there any evidence for non-proportionality in the modelled covariates? Interpret the table and explain your reasoning. (2pt)

```
estat phtest, detail

Test of proportional-hazards assumption
Time:  Time
----------------------------------------------------------------
            |       rho            chi2       df       Prob>chi2
------------+---------------------------------------------------
0b.x1cat    |          .              .        1             .
1.x1cat     |    0.06832           1.38        1        0.2399
2.x1cat     |    0.09127           2.44        1        0.1179
0b.x6cat    |          .              .        1             .
```

```
1.x6cat      |      0.11709          4.01          1          0.0453
hormon       |      0.00780          0.02          1          0.8933
-------------+---------------------------------------------------
global test  |                       7.01          4          0.1351
-----------------------------------------------------------------
```

(f) Based on the previous table and the following plot (Figure 3), how would you expect the hazard ratio for progesterone receptor to vary by time since randomisation? Explain your reasoning. (2pts)

```
estat phtest, plot(1.x6cat)
```

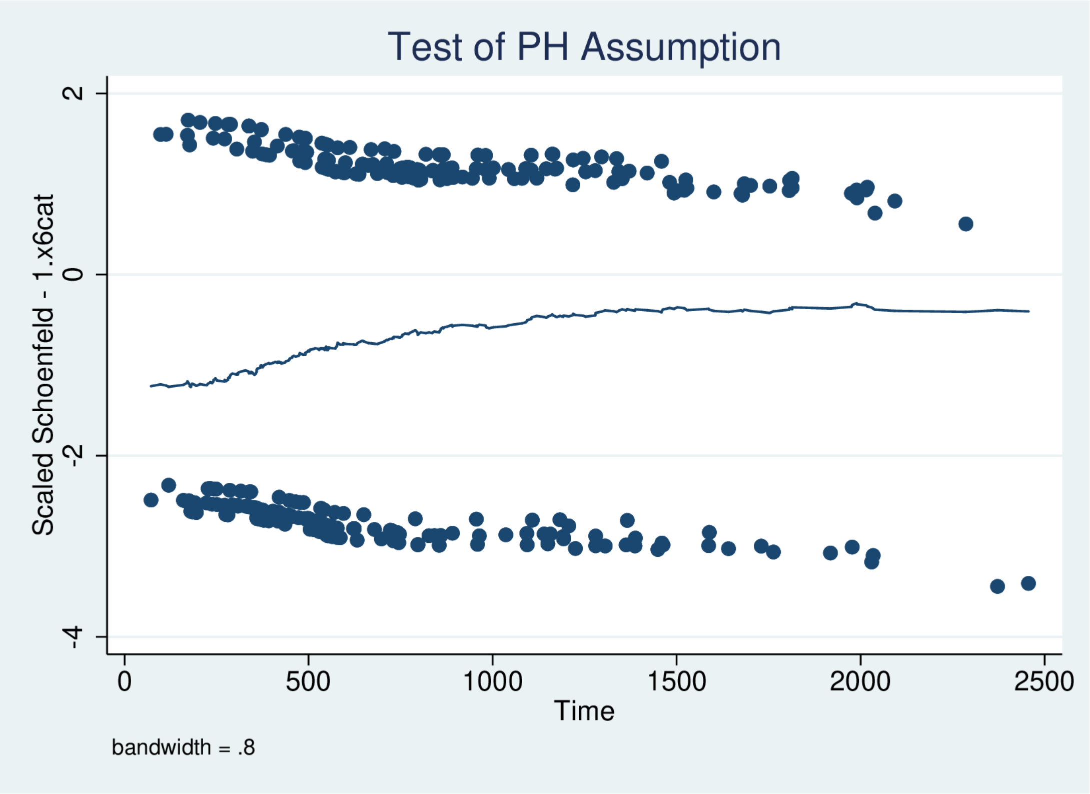

Figure 3: Schoenfeld residual plot for progesterone receptor ≥ 20 fmol, German Breast Cancer Study Group

(g) We now fit a flexible parametric survival model adjusting for **x1cat**, **x6cat** and **hormon** (see the following output). How is this model different to the model in **(d)**? (2pts)

```
stpm2 i.x1cat i.x6cat hormon, baselevels df(4) scale(haz)
est store main_effects

Log likelihood =  -643.9481                     Number of obs     =        684
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
       x1cat |
         0-  |          0  (base)
        45-  |  -.3053002   .1501366    -2.03   0.042    -.5995624   -.0110379
        60-  |  -.1394706   .1641035    -0.85   0.395    -.4611076    .1821665
             |
       x6cat |
         0-  |          0  (base)
        20-  |  -.7626188   .1164153    -6.55   0.000    -.9907887   -.5344489
             |
      hormon |  -.3464509   .1273796    -2.72   0.007    -.5961103   -.0967914
       _rcs1 |   1.532849   .1373159    11.16   0.000     1.263715    1.801983
       _rcs2 |   .4701967   .1341456     3.51   0.000     .2072761    .7331173
       _rcs3 |   .0145754   .0471776     0.31   0.757     -.077891    .1070417
       _rcs4 |  -.0379407   .0179429    -2.11   0.034    -.0731081   -.0027733
       _cons |   -.546419    .142003    -3.85   0.000    -.8247398   -.2680983
------------------------------------------------------------------------------
```

(h) We now fit a model with time-varying effects for progesterone receptor and use a likelihood ratio test to compare the model with time-varying effects with the main effects model in **(g)**. Is there any evidence from the likelihood ratio test for a time-varying effect? (1pt)

```
stpm2 i.x1cat i.x6cat hormon, baselevels df(4) scale(haz) tvc(x6cat) dftvc(2)
est store time_varying
lrtest main_effects time_varying

Log likelihood = -639.81843                     Number of obs     =        684
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
       x1cat |
         0-  |          0  (base)
         45-  |  -.3206187   .1502328    -2.13   0.033    -.6150696   -.0261678
         60-  |  -.1484457   .1643084    -0.90   0.366    -.4704843    .1735929
             |
       x6cat |
         0-  |          0  (base)
         20-  |  -.9279037    .142804    -6.50   0.000    -1.207794    -.648013
             |
      hormon |  -.3500613   .1275366    -2.74   0.006    -.6000283   -.1000942
       _rcs1 |    1.40908   .1490194     9.46   0.000     1.117007    1.701152
       _rcs2 |   .4902992   .1412164     3.47   0.001     .2135202    .7670782
       _rcs3 |   .0193634   .0478939     0.40   0.686    -.0745071    .1132338
       _rcs4 |  -.0424635   .0181817    -2.34   0.020     -.078099    -.006828
 _rcs_x6cat1 |   .3318447   .2159262     1.54   0.124    -.0913627    .7550522
 _rcs_x6cat2 |   .0131242   .1592602     0.08   0.934      -.29902    .3252684
       _cons |  -.4853494   .1432138    -3.39   0.001    -.7660433   -.2046555
------------------------------------------------------------------------------
Likelihood-ratio test                                 LR chi2(2)  =      8.26
(Assumption: main_effects nested in time_varying)     Prob > chi2 =    0.0161
```

(i) From the flexible parametric model with time-varying effects, we plot the time-varying hazard ratio for progesterone receptor ≥ 20 fmol. Carefully interpret the plot in Figure 4. (2pts)

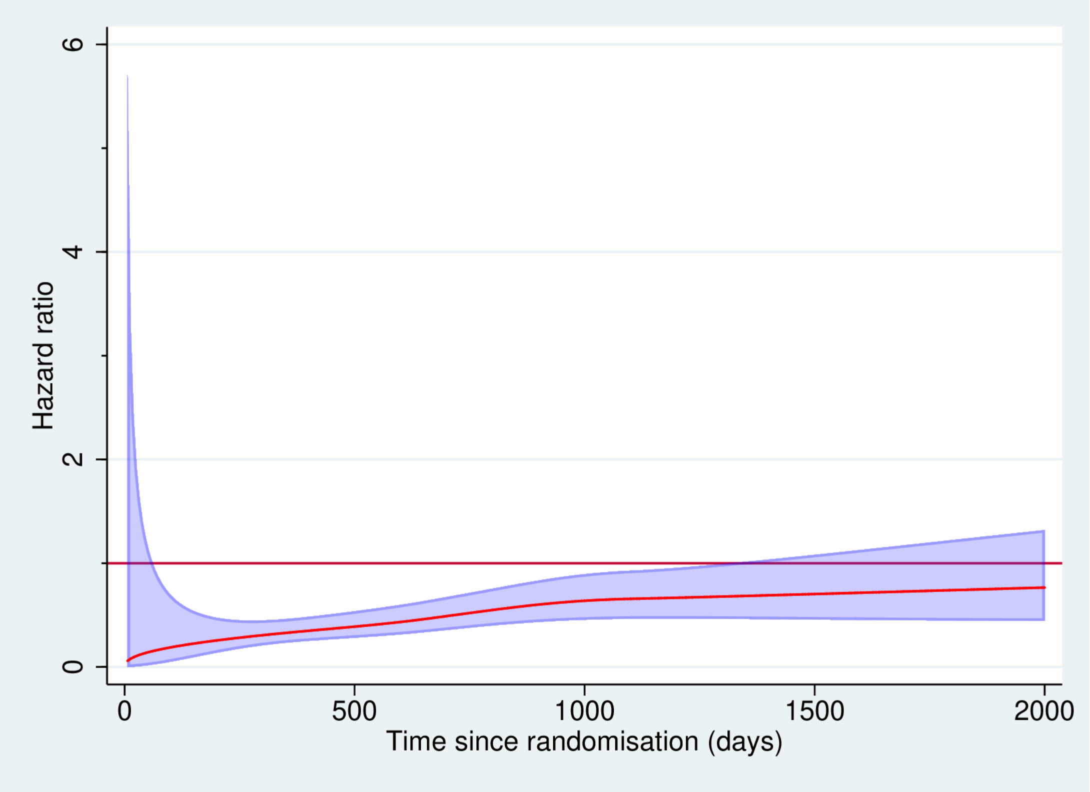

Figure 4: Time-varying hazard ratio for progesterone receptor $\geq$ 20 fmol, German Breast Cancer Study Group

Q3

(a) For a cohort study investigating the onset of diabetes, assume that we have linked general practice (GP) visits with a diabetes quality register and the population register. For each individual diagnosed with diabetes, we assume that diabetes onset happens between the last GP visit when an individual was not diagnosed and the first visit when the individual was diagnosed with diabetes. Also assume that individuals enter the cohort study at different ages and are followed through to death, emigration or 2016-12-01, whichever happens first. Discuss this study design in terms of truncation and censoring. (2pts)

(b) Consider a cancer patient cohort study with two groups defined by cancer stage at diagnosis. The first group has localised or regional spread at cancer diagnosis, and the second group has metastatic spread at cancer diagnosis. For the first group, the five-year survival is 0.8. For the second group, assume we have proportional hazards with a hazard ratio of 2. What is the five-year survival in the second group? Show your working. (1pt)

(c) Compare and contrast (i) a cohort study analysed using Cox regression and (ii) a nested case-control study analysed using conditional logistic regression. How are these two designs and analyses related? What are the advantages and disadvantages of each? (3pts)