Biostat III Examination 2016 Answers

Mark Clements

May 17, 2016

Set-up

```
. global folder 1
. set linesize 80
```

Commentary

In the following answers, the code and full Stata output are provided together with the answers. The full Stata output was not required in the given answers, but is given here to show how the answers were found.

Some brief comments are warranted on presentation. First, when the question asks for specific results, then those results should be presented separately in text, rather than only presenting the output from the statistical package. Second, the choice of non-proportional fonts makes it difficult to read output from the statistical package. Third, using colours in the graphics makes it difficult to discern which line is which in black-and-white printout. I suggest that using `scheme(s2mono)` would be useful for graphics in Stata.

Part 1

Question 1

We read in the dataset:

```
. import delimited "http://biostat3.net/download/exams/2016/$folder/incidence.c
> sv", clear
(6 vars, 360 obs)
. egen agecat = cut(age), at(40, 50, 60, 70, 80, 90)
```

We then fit a Poisson regression with the number of lung cancer cases at the outcome (first argument), with the person-time of exposure as the `exposure` option. We include attained `age` as a linear, continuous effect in each model.

```
. poisson lc sex age, exposure(pt) nolog irr

Poisson regression                              Number of obs   =        360
                                                LR chi2(2)      =     493.81
                                                Prob > chi2     =     0.0000
Log likelihood = -837.39175                     Pseudo R2       =     0.2277

------------------------------------------------------------------------------
          lc |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   2.303417   .2217275     8.67   0.000     1.907373    2.781694
         age |   1.090729   .0045982    20.60   0.000     1.081754    1.099779
       _cons |   2.13e-06   5.87e-07   -47.32   0.000     1.24e-06    3.65e-06
      ln(pt) |          1  (exposure)
------------------------------------------------------------------------------
```

```
------------------------------------------------------------------------------
. poisson lc smoking age, exposure(pt) nolog irr

Poisson regression                                Number of obs   =        360
                                                  LR chi2(2)      =    1172.97
                                                  Prob > chi2     =     0.0000
Log likelihood = -497.81498                       Pseudo R2       =     0.5409

------------------------------------------------------------------------------
          lc |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     smoking |   14.59096   1.666623    23.47   0.000     11.66425    18.25201
         age |   1.096176   .0046727    21.54   0.000     1.087056    1.105373
       _cons |   6.24e-07   1.80e-07   -49.56   0.000     3.55e-07    1.10e-06
      ln(pt) |          1  (exposure)
------------------------------------------------------------------------------
. poisson lc asbestos age, exposure(pt) nolog irr

Poisson regression                                Number of obs   =        360
                                                  LR chi2(2)      =     524.76
                                                  Prob > chi2     =     0.0000
Log likelihood = -821.91845                       Pseudo R2       =     0.2420

------------------------------------------------------------------------------
          lc |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    asbestos |   3.679055   .3976635    12.05   0.000     2.976674    4.547172
         age |   1.089761   .0045805    20.45   0.000      1.08082    1.098776
       _cons |   3.06e-06   8.16e-07   -47.61   0.000     1.81e-06    5.16e-06
      ln(pt) |          1  (exposure)
------------------------------------------------------------------------------
```

The age-adjusted incidence rate ratio for sex is 2.30 (95% confidence interval (CI): 1.91, 2.78). This association is highly significant ($p < 0.001$).

The age-adjusted incidence rate ratio for smoking is 14.59 (95% confidence interval (CI): 11.66, 18.25). This association is highly significant ($p < 0.001$). The age-adjusted incidence rate ratio for asbestos is 3.68 (95% confidence interval (CI): 2.98, 4.55). This association is highly significant ($p < 0.001$).

We could have adjusted for attained age in several other ways, including quintiles or splines. To investigate this, we first use quintiles with sex:

```
. xtile ageQ5 = age, nquantiles(5)
. poisson lc sex i.ageQ5, exposure(pt) nolog irr base

Poisson regression                                Number of obs   =        360
                                                  LR chi2(5)      =     463.73
                                                  Prob > chi2     =     0.0000
Log likelihood = -852.43355                       Pseudo R2       =     0.2138

------------------------------------------------------------------------------
          lc |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   2.282888   .2197354     8.58   0.000       1.8904    2.756865
             |
       ageQ5 |
          1  |          1  (base)
          2  |   2.838344    .479085     6.18   0.000     2.038874    3.951296
          3  |   6.696689    1.06818    11.92   0.000      4.89876    9.154489
          4  |   12.24698   1.997027    15.36   0.000     8.896714    16.85886
```

```
          5  |   17.04721   3.612093    13.38   0.000     11.25367    25.82334
             |
       _cons |   .0000798   .0000123   -61.29   0.000      .000059    .0001079
      ln(pt) |          1  (exposure)
------------------------------------------------------------------------------
```

This shows a very similar point estimate and standard errors to modelling attained age as a linear, continuous effect. We also investigate using restricted cubic splines:

```
. mkspline ageSpline = age, cubic nknots(4)
. poisson lc sex ageSpline*, exposure(pt) nolog irr base

Poisson regression                                Number of obs   =        360
                                                  LR chi2(4)      =     504.80
                                                  Prob > chi2     =     0.0000
Log likelihood = -831.89916                       Pseudo R2       =     0.2328

------------------------------------------------------------------------------
          lc |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |    2.29398   .2208203     8.63   0.000     1.899557    2.770301
  ageSpline1 |   1.140294   .0247602     6.05   0.000     1.092783     1.18987
  ageSpline2 |   .9210111   .0576881    -1.31   0.189     .8146092    1.041311
  ageSpline3 |   1.131983   .1996173     0.70   0.482      .801192     1.59935
       _cons |   2.30e-07   2.46e-07   -14.30   0.000     2.83e-08    1.87e-06
      ln(pt) |          1  (exposure)
------------------------------------------------------------------------------
```

Again, this shows a very similar point estimate and standard errors to modelling attained age as a linear, continuous effect. I accepted answers using any of quintiles, linear/continuous age, splines or similar functional forms.

In summary, lung cancer incidence is associated with age, sex, asbestos exposure and current smoking exposure.

Question 2

We now adjust for age, sex, smoking exposure and asbestos exposure in the same model.

```
. poisson lc age sex smoking asbestos, exposure(pt) nolog irr

Poisson regression                                Number of obs   =        360
                                                  LR chi2(4)      =    1299.99
                                                  Prob > chi2     =     0.0000
Log likelihood = -434.30591                       Pseudo R2       =     0.5995

------------------------------------------------------------------------------
          lc |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.097531     .00469    21.78   0.000     1.088377    1.106761
         sex |   1.567733   .1529406     4.61   0.000     1.294891    1.898066
     smoking |   13.66723   1.568787    22.78   0.000      10.9138    17.11533
    asbestos |   3.271022   .3565175    10.87   0.000     2.641854     4.05003
       _cons |   3.92e-07   1.17e-07   -49.66   0.000     2.19e-07    7.02e-07
      ln(pt) |          1  (exposure)
------------------------------------------------------------------------------
. est store ModelA
```

This shows clearly that each of attained age, sex, smoking and asbestos exposure are significantly associated with lung cancer incidence ($p < 0.001$ for all adjusted effects). The adjusted rate ratio (RR)

for age was 1.098 (95% CI: 1.088, 1.107) per year of age, indicating a rapid rise with increasing age. Males have higher rates of disease even after adjustment for other covariates (RR=1.57, 95% CI: 1.29, 1.90). Smoking is strongly associated with lung cancer incidence (RR=13.67, 95% CI: 10.91, 17.12). Finally, asbestos exposure has a rate ratio of 3.27 (95% CI: 2.64, 4.05).

Empirical evidence for confounding can be assessed in several ways. First, we can assess whether exposure to smoking and asbestos are associated:

```
. tab smoking asbestos [aw=pt], row

+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+

           |       asbestos
   smoking |         0          1 |     Total
-----------+----------------------+----------
         0 | 253.95541  19.564223 | 273.51963
           |     92.85       7.15 |    100.00
-----------+----------------------+----------
         1 | 79.645205  6.8351613 | 86.480366
           |     92.10       7.90 |    100.00
-----------+----------------------+----------
     Total | 333.60062  26.399385 |       360
           |     92.67       7.33 |    100.00
```

We see that exposure to asbestos is reasonably similar between never smokers (7.2%) and current smokers (7.9%), suggesting that asbestos and smoking exposure are not associated, and therefore are unlikely to be confounded. We are not able to undertake a formal statistical test with these weighted data.

Second, we can assess whether the estimated associations between lung cancer incidence and each of smoking and asbestos change after an adjustment for other covariates.

Comparing the linear age-adjusted model with a model with attained age, smoking and asbestos, we see that the rate ratio for asbestos changed from 3.70 to 3.51 (5% reduction), and the rate ratio for smoking changed from 14.59 to 14.42 (1% reduction). Again, there is limited evidence for confounding between smoking and asbestos.

Question 3

(a)

A regression model formula is

$$\begin{aligned}\log(\lambda(t|x)) \quad = \quad & \beta_0 + \beta_1 \text{age} + \beta_2 I(\text{sex} = 1) + \beta_3 I(\text{smoking} = 1) + \beta_4 I(\text{asbestos} = 1) + \\ & \beta_5 I(\text{smoking} = 1 \ \& \ \text{asbestos} = 1)\end{aligned}$$

where $\lambda(t|x)$ is the rate at attained age t given covariates x (including sex, smoking and asbestos), with coefficients $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ and β_5, and $I(test)$ is 1 if the test is true and 0 if the test is false.

(b)

We now fit the interaction model:

```
. poisson lc age sex smoking##asbestos, exposure(pt) nolog irr

Poisson regression                              Number of obs   =        360
                                                LR chi2(5)      =    1305.70
                                                Prob > chi2     =     0.0000
```

```
Log likelihood = -431.44866                     Pseudo R2       =     0.6021

------------------------------------------------------------------------------
          lc |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    1.09749   .0046902    21.77   0.000     1.088335    1.106721
         sex |   1.563397   .1523218     4.59   0.000     1.291626    1.892352
   1.smoking |   15.95602   2.166946    20.40   0.000     12.22714    20.82207
  1.asbestos |   5.354885   1.196287     7.51   0.000     3.456136    8.296779
             |
     smoking#|
    asbestos |
        1 1  |   .5353407   .1366072    -2.45   0.014     .3246551    .8827512
             |
       _cons |   3.47e-07   1.05e-07   -48.99   0.000     1.92e-07    6.29e-07
      ln(pt) |          1  (exposure)
------------------------------------------------------------------------------
. est store ModelB
. lrtest ModelA ModelB

Likelihood-ratio test                                 LR chi2(1)  =      5.71
(Assumption: ModelA nested in ModelB)                 Prob > chi2 =    0.0168
```

Comparing Model A with Model B, we see that there is good evidence for a statistical interaction on a multiplicative scale. First, we note that the Wald test for the interaction term has a p-value of 0.014. Second, we see that the likelihood ratio test is also highly significant, with $p = 0.017$. We can re-express the rate ratios for different combinations of smoking and asbestos:

```
. poisson lc age sex smoking#asbestos, exposure(pt) nolog irr

Poisson regression                              Number of obs   =        360
                                                LR chi2(5)      =    1305.70
                                                Prob > chi2     =     0.0000
Log likelihood = -431.44866                     Pseudo R2       =     0.6021

------------------------------------------------------------------------------
          lc |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    1.09749   .0046902    21.77   0.000     1.088335    1.106721
         sex |   1.563397   .1523218     4.59   0.000     1.291626    1.892352
             |
     smoking#|
    asbestos |
        0 1  |   5.354885   1.196287     7.51   0.000     3.456136    8.296779
        1 0  |   15.95602   2.166946    20.40   0.000     12.22714    20.82207
        1 1  |   45.74092   7.678061    22.77   0.000     32.91716    63.56051
             |
       _cons |   3.47e-07   1.05e-07   -48.99   0.000     1.92e-07    6.29e-07
      ln(pt) |          1  (exposure)
------------------------------------------------------------------------------
```

This shows that lung cancer incidence rate ratio for exposure to both asbestos and smoking is 45.74 (95% CI: 32.92, 63,56) compared with no exposure to both risk factors. However, the effect of both risk factors is significantly less than multiplicative.

(c)

From Model B, we can calculate the incidence rate for a males aged 62 years who has been exposed to asbestos and is a current smoker using several approaches. By hand, we can calculate the rate as

$3.47e-07 \times 1.09749^{62} \times 1.563397 \times 45.74092 \approx 0.0079$ per person-year. To calculate the confidence interval, we need to take account of the covariance terms, which is best done using tools provided by each statistical package. Using the `lincom` command:

```
. quietly poisson lc age sex smoking##asbestos, exposure(pt) nolog irr
. lincom sex + 1.smoking + 1.asbestos + 1.smoking#1.asbestos + 62*age + _cons,
> irr

 ( 1)  62*[lc]age + [lc]sex + [lc]1.smoking + [lc]1.asbestos +
       [lc]1.smoking#1.asbestos + [lc]_cons = 0

------------------------------------------------------------------------------
          lc |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .0079392   .0008947   -42.91   0.000     .0063659    .0099015
------------------------------------------------------------------------------
```

This shows that the incidence rate is 7.94 (95% CI: 6.37, 9.90) per 1000 person-years.

We can do the same analysis using the `margins` and `predict` commands.

Part 2

Question 4

We read in the data using the following:

```
. display "Folder = $folder"
Folder = 1
. import delimited "http://biostat3.net/download/exams/2016/$folder/survival.cs
> v", clear
(8 vars, 496 obs)
```

(a)

This question is equivalent to completing *Table 1* for a randomised controlled trial to assess whether randomisation led to balanced covariates. We use simple tests to assess whether treatment assignment varies substantially by age at diagnosis, sex, smoking exposure and asbestos exposure.

For age at diagnosis, we can use either a t-test or a non-parametric test:

```
. ttest age, by(tx)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     237    63.19049    .6275862    9.661567     61.9541    64.42687
       1 |     259    62.27185    .6214458    10.00122     61.0481     63.4956
---------+--------------------------------------------------------------------
combined |     496     62.7108     .441883    9.841202     61.8426    63.57899
---------+--------------------------------------------------------------------
    diff |            .9186377    .8845665               -.8193388    2.656614
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   1.0385
Ho: diff = 0                                     degrees of freedom =      494

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.8502         Pr(|T| > |t|) = 0.2995          Pr(T > t) = 0.1498
. ranksum age, by(tx)
```

```
Two-sample Wilcoxon rank-sum (Mann-Whitney) test

          tx |      obs    rank sum    expected
-------------+---------------------------------
           0 |      237       60434     58894.5
           1 |      259       62822     64361.5
-------------+---------------------------------
    combined |      496      123256      123256

unadjusted variance  2542279.25
adjustment for ties        0.00
                     ----------
adjusted variance    2542279.25

Ho: age(tx==0) = age(tx==1)
             z =   0.966
    Prob > |z| =   0.3343
```

We find some evidence that age differs by treatment modality ($p = 0.028$ for the t-test and $p = 0.050$ for the Wilcoxon test), where the lung cancer patients randomised to conventional therapy are slightly older than patients randomised to chemotherapy+radiotherapy. For the other variables:

```
. tab tx sex, chi row

+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+

           |          sex
        tx |         0          1 |     Total
-----------+----------------------+----------
         0 |        88        149 |       237
           |     37.13      62.87 |    100.00
-----------+----------------------+----------
         1 |        71        188 |       259
           |     27.41      72.59 |    100.00
-----------+----------------------+----------
     Total |       159        337 |       496
           |     32.06      67.94 |    100.00

          Pearson chi2(1) =   5.3657   Pr = 0.021
. tab tx smoking, chi row

+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+

           |        smoking
        tx |         0          1 |     Total
-----------+----------------------+----------
         0 |        46        191 |       237
           |     19.41      80.59 |    100.00
-----------+----------------------+----------
```

```
         1 |        49        210 |       259
           |     18.92      81.08 |    100.00
-----------+----------------------+----------
     Total |        95        401 |       496
           |     19.15      80.85 |    100.00

          Pearson chi2(1) =   0.0192   Pr = 0.890
. tab tx asbestos, chi row

+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+

           |       asbestos
        tx |         0          1 |     Total
-----------+----------------------+----------
         0 |       192         45 |       237
           |     81.01      18.99 |    100.00
-----------+----------------------+----------
         1 |       194         65 |       259
           |     74.90      25.10 |    100.00
-----------+----------------------+----------
     Total |       386        110 |       496
           |     77.82      22.18 |    100.00

          Pearson chi2(1) =   2.6762   Pr = 0.102
```

We again find some evidence that randomisation varied by sex, where there are the proportion of males n those randomised to conventional is lower than for those randomised to chemotherapy+radiotherapy ($p = 0.002$). There is no evidence for imbalance by smoking and asbestos exposure.

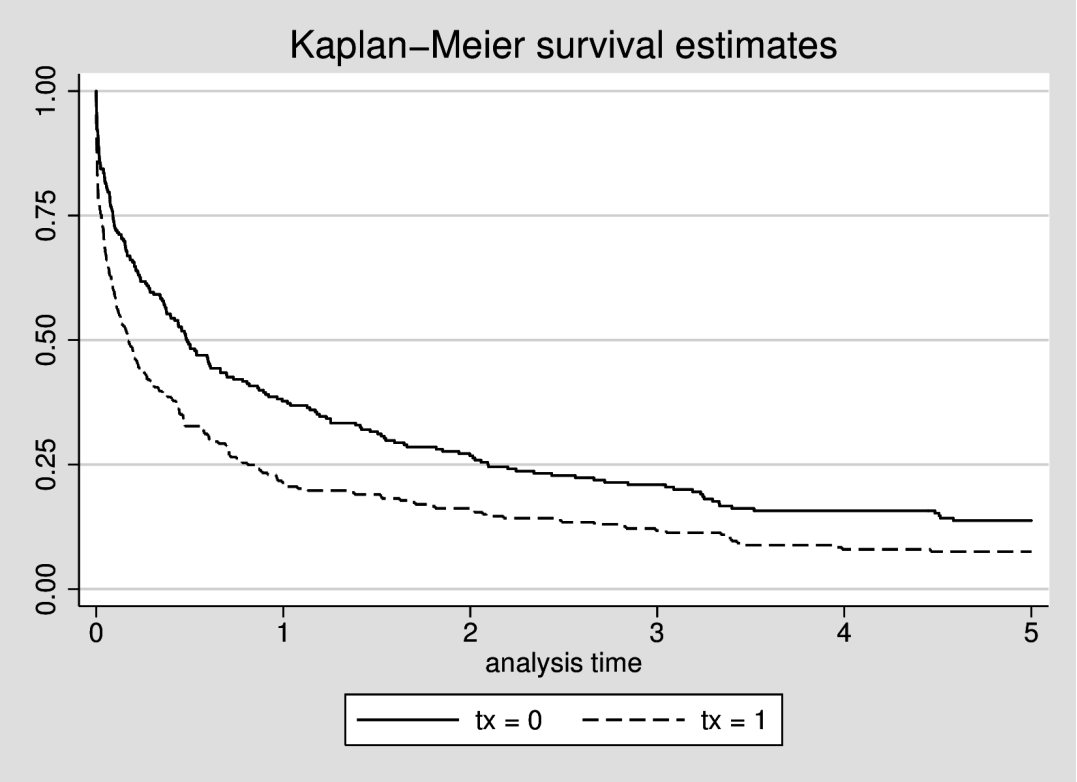
In summary, further analyses are potentially confounded by age and sex and we should consider adjusting for those variables in the survival analysis.

(b)

We `stset` the data using time since diagnosis as the primary time scale and then plot the Kaplan-Meier curves

```
. stset tsurv, failure(event) id(id)

                id:  id
     failure event:  event != 0 & event < .
obs. time interval:  (tsurv[_n-1], tsurv]
 exit on or before:  failure

--------------------------------------------------------------------------------
      496  total observations
        0  exclusions
--------------------------------------------------------------------------------
      496  observations remaining, representing
      496  subjects
      435  failures in single-failure-per-subject data
 530.7275  total analysis time at risk and under observation
                                             at risk from t =         0
                                  earliest observed entry t =         0
                                       last observed exit t =         5
```

```
. sts graph, by(tx) name(km1, replace) scheme(s2mono)

         failure _d:  event
   analysis time _t:  tsurv
                 id:  id
. graph export exam_2016_km1.eps, name(km1) replace
(file exam_2016_km1.eps written in EPS format)
. * the following line is only needed on Linux
. !! convert -density 300 exam_2016_km1.eps exam_2016_km1.png
. sts test tx

         failure _d:  event
   analysis time _t:  tsurv
                 id:  id


Log-rank test for equality of survivor functions
------------------------------------------------

      |   Events         Events
tx    |  observed       expected
------+-------------------------
0     |       198         239.65
1     |       237         195.35
------+-------------------------
Total |       435         435.00

           chi2(1) =      16.23
           Pr>chi2 =     0.0001
. sts list, by(tx) at(1 2 3 4 5)

         failure _d:  event
   analysis time _t:  tsurv
                 id:  id

             Beg.                      Survivor      Std.
    Time     Total     Fail             Function     Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
tx=0
       1        87      145               0.3774    0.0319     0.3151    0.4395
       2        62       25               0.2677    0.0292     0.2122    0.3261
       3        46       13               0.2095    0.0270     0.1593    0.2645
       4        34       11               0.1571    0.0244     0.1129    0.2081
       5        28        4               0.1375    0.0233     0.0959    0.1866
tx=1
       1        55      203               0.2136    0.0256     0.1658    0.2656
       2        42       13               0.1622    0.0231     0.1200    0.2100
       3        29       11               0.1174    0.0203     0.0813    0.1606
       4        20        9               0.0797    0.0172     0.0502    0.1177
       5        17        1               0.0752    0.0168     0.0467    0.1126
-------------------------------------------------------------------------------
Note:  survivor function is calculated over full data and evaluated at
       indicated times; it is not calculated from aggregates shown at left.
```



The Kaplan-Meier curves show that survival is poor for lung cancer patients, with fewer than 25% of patients surviving to 5 years. We also see that treatment with chemotherapy+radiotherapy leads to more deaths soon after diagnosis. It is unclear whether the rates are different after one year.

Although not specifically asked for, we also (i) used the log-rank test to compare the curves, finding strong evidence for a difference ($p = 0.0001$) and (ii) estimated survival to five years, where 14% (95% CI: 10, 19) survived for those on conventional treatment and 8% (95% CI: 5, 11) survived for those on chemotherapy+radiotherapy.

Question 5

Based on Question 4 (a), we first investigated whether age and sex were associated with survival and hence would be potential confounders:

```
. stcox tx sex age, nolog

        failure _d:  event
  analysis time _t:  tsurv
                id:  id

Cox regression -- no ties

No. of subjects =          496                     Number of obs    =       496
No. of failures =          435
Time at risk    =  530.7275306
                                                   LR chi2(3)       =     16.68
Log likelihood  =   -2370.0399                     Prob > chi2      =    0.0008

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          tx |   1.474455   .1447368     3.96   0.000     1.216397    1.787261
         sex |   .9460891   .0975808    -0.54   0.591     .7729263    1.158047
         age |   .9974118   .0048443    -0.53   0.594     .9879621    1.006952
------------------------------------------------------------------------------
```

```
. stcox tx sex, nolog

        failure _d:  event
  analysis time _t:  tsurv
                id:  id

Cox regression -- no ties

No. of subjects =          496                     Number of obs   =       496
No. of failures =          435
Time at risk    =  530.7275306
                                                   LR chi2(2)      =     16.40
Log likelihood  =   -2370.1819                     Prob > chi2     =    0.0003

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          tx |   1.482879   .1446844     4.04   0.000     1.224768    1.795384
         sex |   .9466926   .0976307    -0.53   0.595      .773439    1.158756
------------------------------------------------------------------------------
. stcox tx age, nolog

        failure _d:  event
  analysis time _t:  tsurv
                id:  id

Cox regression -- no ties

No. of subjects =          496                     Number of obs   =       496
No. of failures =          435
Time at risk    =  530.7275306
                                                   LR chi2(2)      =     16.40
Log likelihood  =   -2370.1833                     Prob > chi2     =    0.0003

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          tx |    1.46388   .1423281     3.92   0.000     1.209891    1.771189
         age |   .9974397   .0048479    -0.53   0.598     .9879832    1.006987
------------------------------------------------------------------------------
. stcox tx, nolog

        failure _d:  event
  analysis time _t:  tsurv
                id:  id

Cox regression -- no ties

No. of subjects =          496                     Number of obs   =       496
No. of failures =          435
Time at risk    =  530.7275306
                                                   LR chi2(1)      =     16.12
Log likelihood  =   -2370.3221                     Prob > chi2     =    0.0001

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          tx |   1.472317   .1422442     4.00   0.000     1.218329    1.779254
```

Adjusting for treatment modality, there is no evidence that either age or sex are associated with survival, with Wald test p-values of 0.60 for both age and sex. Furthermore, fitting a Cox regression models with and without age and sex suggest that the effect of treatment modality is insensitive to inclusion of age and sex in the model. The hazard ratio for chemotherapy+radiotherapy compared with conventional therapy is 1.47 (95% CI: 1.22, 1.78), suggesting that the average hazard ratio for chemotherapy+radiotherapy is high over the five-year period.

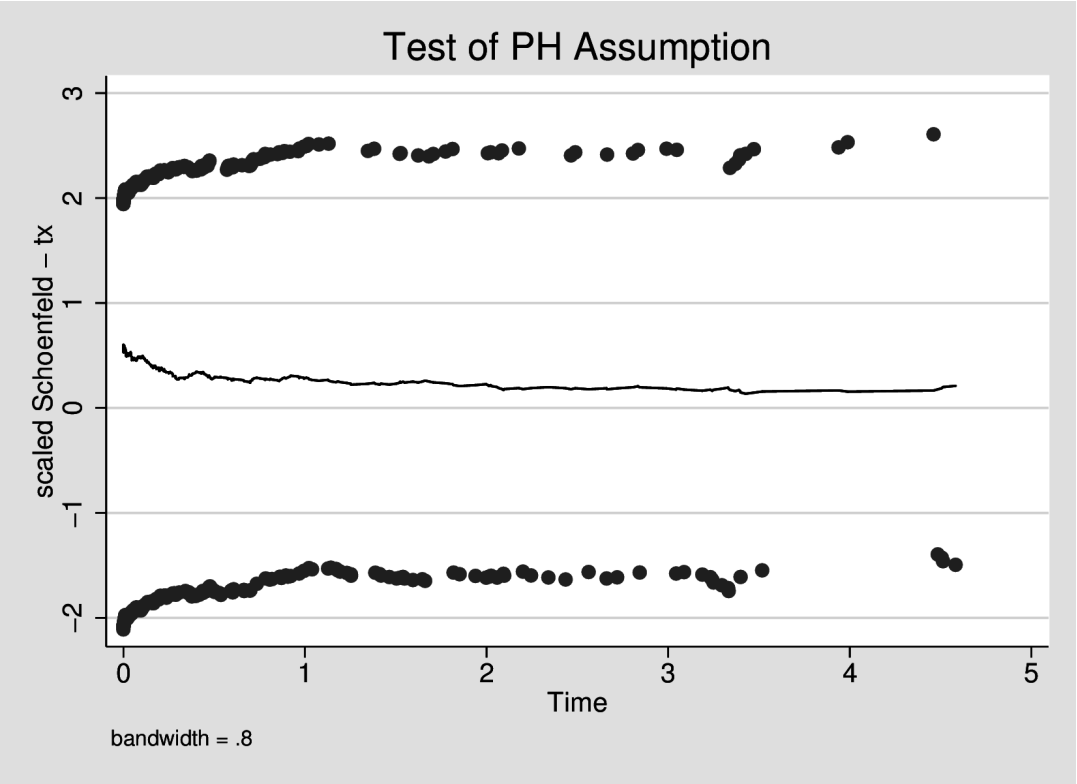
For the time scale, we have initially used time since cancer diagnosis. There is a strong association between time since diagnosis and survival, suggesting that this is the best choice of primary time scale. Moreover, there is a suggestion of non-proportional hazards, with a higher rate ratio in the first year than for the later years. We could investigate using attained age as the primary time scale, but then we would need to finely model for the time since diagnosis, which would require modelling two time scales. For simplicity, we propose using time since diagnosis as the primary time scale.

Question 6

(i)

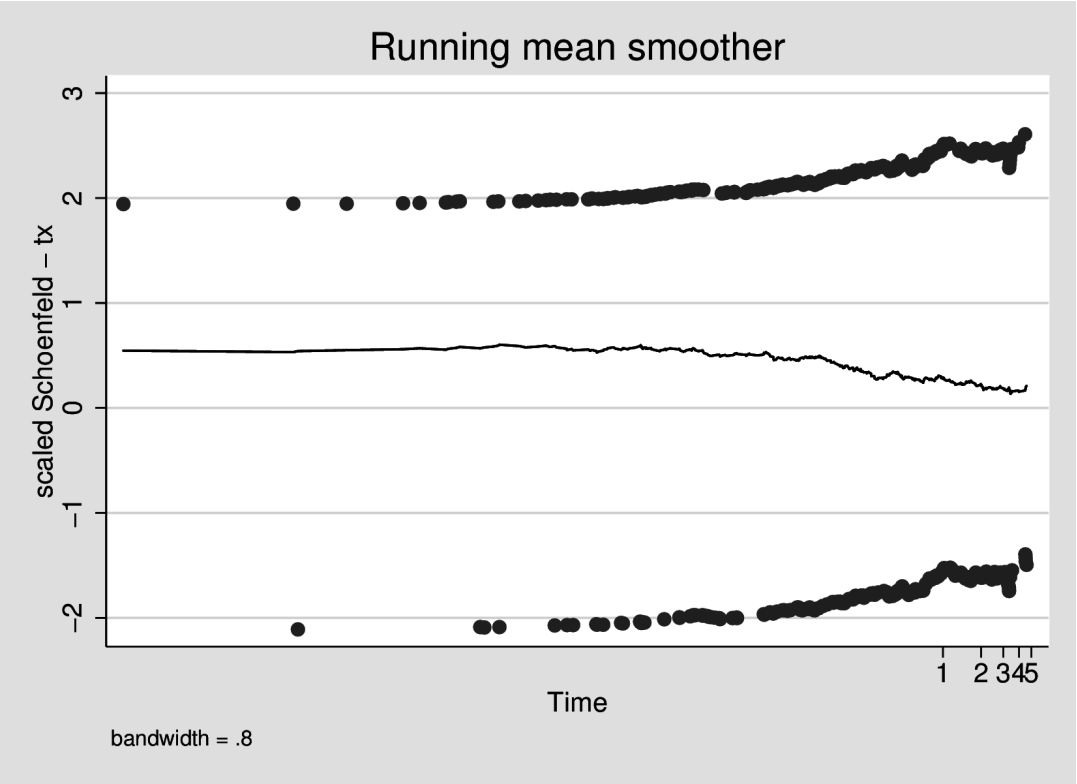
For an analysis of scaled Schoenfeld residuals, we use:

```
. estat phtest, detail

      Test of proportional-hazards assumption

      Time:  Time
      ----------------------------------------------------------------
                  |       rho            chi2       df       Prob>chi2
      ------------+---------------------------------------------------
      tx          |     -0.07976         2.71        1         0.0999
      ------------+---------------------------------------------------
      global test |                      2.71        1         0.0999
      ----------------------------------------------------------------
. estat phtest, plot(tx) name(phtest, replace) scheme(s2mono)
. graph export exam_2016_phtest.eps, name(phtest) replace
(file exam_2016_phtest.eps written in EPS format)
. * the following line is only needed on Linux
. !! convert -density 300 exam_2016_phtest.eps exam_2016_phtest.png
```



This shows that there is some evidence ($p = 0.10$) that the hazard ratio decreases with increasing time since diagnosis: the scaled residuals and linear time have a correlation of -0.08. From the plot of the scaled residuals and time, we see the running mean smoother dips early in the follow-up period and then is flat or very slightly declining. Given the number of events that are early in the period, we could also test using a log-transformation for time since diagnosis:

```
. estat phtest, detail log

      Test of proportional-hazards assumption

      Time:  Log(t)
      ------------------------------------------------------------------
                  |       rho            chi2       df       Prob>chi2
      ------------+-----------------------------------------------------
      tx          |     -0.10739         4.91        1         0.0267
      ------------+-----------------------------------------------------
      global test |                      4.91        1         0.0267
      ------------------------------------------------------------------
. estat phtest, log plot(tx) name(phtestlog, replace) scheme(s2mono)
. graph export exam_2016_phtestlog.eps, name(phtestlog) replace
(file exam_2016_phtestlog.eps written in EPS format)
. * the following line is only needed on Linux
. !! convert -density 300 exam_2016_phtestlog.eps exam_2016_phtestlog.png
```



This transformation indicates a stronger association than that for untransformed time ($p = 0.027$), with evidence for linearity in the hazard ratio on the log-time scale.

(ii)

We can test for piecewise-constant hazard ratios by splitting by time and fitting for an interaction. In the following, the "c" prefix indicates a continuous variable, while the "i" prefix indicates a factor variable.

```
. quietly import delimited "http://biostat3.net/download/exams/2016/$folder/sur
> vival.csv", clear
. quietly stset tsurv, fail(event) id(id)
. stsplit timeband, at(0, 1, max)
(140 observations (episodes) created)
. stcox sex i.tx##i.timeband, nolog

        failure _d:  event
  analysis time _t:  tsurv
                id:  id

Cox regression -- no ties

No. of subjects =          496                     Number of obs    =       636
No. of failures =          435
Time at risk    =  530.7275306
                                                   LR chi2(3)       =     20.57
Log likelihood  =   -2368.0965                     Prob > chi2      =    0.0001

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .9535614   .0984221    -0.46   0.645     .7789179    1.167362
       1.tx |   1.636955   .1794447     4.50   0.000     1.320465      2.0293
  1.timeband |   7.392735          .        .       .            .           .
```

```
             |
 tx#timeband |
       1 1   |   .6086022   .1494051    -2.02   0.043     .3761602    .9846777
------------------------------------------------------------------------------
. stcox tx sex c.tx#c.timeband, nolog

        failure _d:  event
  analysis time _t:  tsurv
                id:  id

Cox regression -- no ties

No. of subjects =          496                     Number of obs   =       636
No. of failures =          435
Time at risk    =  530.7275306
                                                   LR chi2(3)      =     20.57
Log likelihood  =   -2368.0965                     Prob > chi2     =    0.0001

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          tx |   1.636955   .1794447     4.50   0.000     1.320465      2.0293
         sex |   .9535614   .0984221    -0.46   0.645     .7789179    1.167362
             |
        c.tx#|
  c.timeband |   .6086022   .1494051    -2.02   0.043     .3761602    .9846777
------------------------------------------------------------------------------
. stcox c.tx#i.timeband, nolog

        failure _d:  event
  analysis time _t:  tsurv
                id:  id

Cox regression -- no ties

No. of subjects =          496                     Number of obs   =       636
No. of failures =          435
Time at risk    =  530.7275306
                                                   LR chi2(2)      =     20.36
Log likelihood  =    -2368.202                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
           _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
timeband#c.tx |
           0  |   1.628084   .1774136     4.47   0.000     1.314985    2.015732
           1  |   .9871128   .2169872    -0.06   0.953      .641587    1.518721
------------------------------------------------------------------------------
```

This model provides some evidence that the hazard ratio is time-dependent ($p = 0.04$). The hazard ratio in the first year is 1.63 (95% CI: 1.31, 2.02), while the hazard ratio in the second year is 0.99 (95% CI: 0.64, 1.52).

(iii)

We can re-fit the model in (ii) using Stata `stcox`'s `tvc` and `texp` options:

```
. stcox tx, nolog tvc(tx) texp(_t>=1)
```

```
        failure _d:  event
  analysis time _t:  tsurv
                id:  id


Cox regression -- no ties

No. of subjects =          496                     Number of obs   =       636
No. of failures =          435
Time at risk    =  530.7275306
                                                   LR chi2(2)      =     20.36
Log likelihood  =    -2368.202                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
main         |
          tx |   1.628084   .1774136     4.47   0.000     1.314985    2.015732
-------------+----------------------------------------------------------------
tvc          |
          tx |   .6063033   .1487552    -2.04   0.041     .3748432    .9806868
------------------------------------------------------------------------------
Note: variables in tvc equation interacted with _t>=1
```

Again, we find some evidence for a time-dependent hazard ratio. We can model for a time-dependent hazard ratio that depends on time or log(time):

```
. stcox tx, nolog tvc(tx) texp(_t)

        failure _d:  event
  analysis time _t:  tsurv
                id:  id


Cox regression -- no ties

No. of subjects =          496                     Number of obs   =       636
No. of failures =          435
Time at risk    =  530.7275306
                                                   LR chi2(2)      =     18.84
Log likelihood  =    -2368.961                     Prob > chi2     =    0.0001

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
main         |
          tx |   1.639451    .192211     4.22   0.000     1.302875    2.062976
-------------+----------------------------------------------------------------
tvc          |
          tx |   .8468213   .0864529    -1.63   0.103     .6932529    1.034408
------------------------------------------------------------------------------
Note: variables in tvc equation interacted with _t
```

The interpretation of this model is as follows: the hazard ratio at time 0 is 1.64 (95% CI: 1.30, 2.06); for each year since diagnosis, the rate tends to decrease by 1-0.85=15% (RR=0.85, 95% CI: 0.69, 1.03), although this trend is not significant ($p = 0.10$, as per the Schoenfeld test). We could also model for time-dependence using log-time:

```
. stcox tx, nolog tvc(tx) texp(log(_t))
```

```
        failure _d:  event
  analysis time _t:  tsurv
                id:  id


Cox regression -- no ties

No. of subjects =          496                     Number of obs   =       636
No. of failures =          435
Time at risk    =  530.7275306
                                                   LR chi2(2)      =     21.28
Log likelihood  =   -2367.7426                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
main         |
          tx |   1.227843   .1550932     1.62   0.104     .9585718    1.572755
-------------+----------------------------------------------------------------
tvc          |
          tx |   .9151138    .036646    -2.22   0.027     .8460352    .9898325
------------------------------------------------------------------------------
Note: variables in tvc equation interacted with log(_t)
```

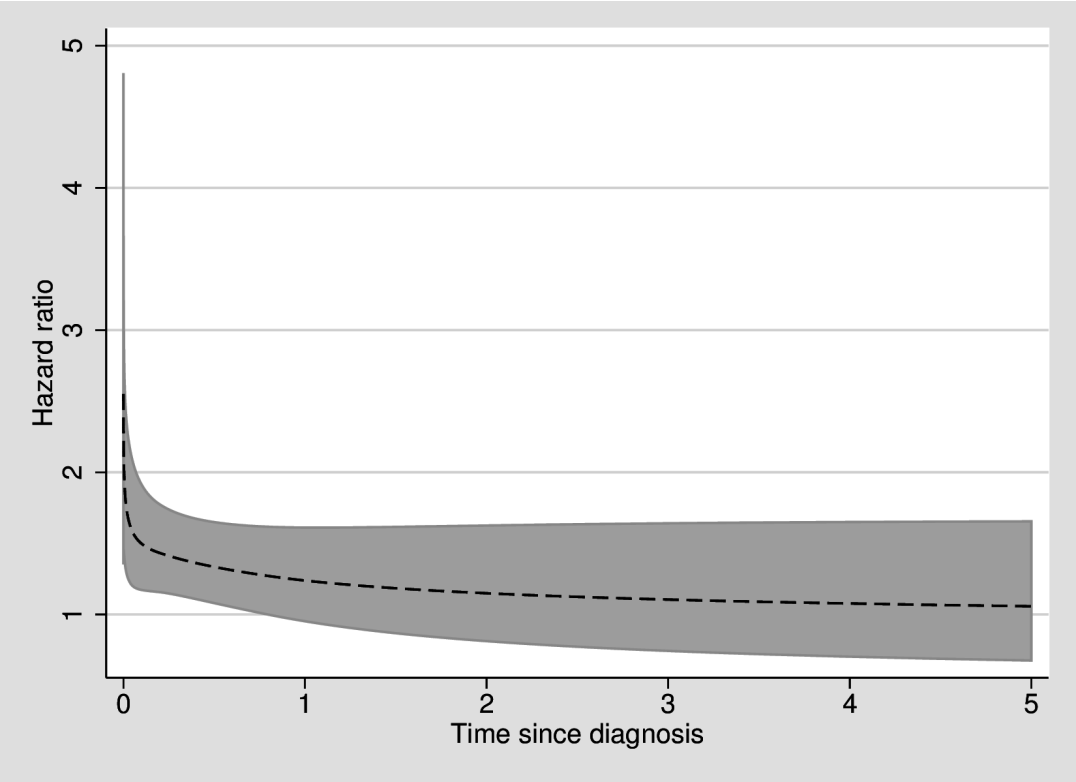
This model can be interpreted as follows: at time equals 1, the hazard ratio is 1.23 (95% CI: 0.96, 1.57; $p = 0.10$); for each unit increase in log(time), the hazard ratio is multiplied by a factor of 0.92 (95% CI: 0.85, 0.99; $p = 0.027$).

(iv)

Using `stpm2` with time-dependent hazard ratios, we use a low-dimensional natural spline for the time-dependent effect. We use a Wald test to check for time-dependence and plot the time-dependent hazard ratio:

```
. stpm2 tx, df(4) scale(hazard) nolog eform tvc(tx) dftvc(2)
note: delayed entry models are being fitted

Log likelihood = -1124.2991                        Number of obs   =       636

------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
          tx |   1.672689   .1897623     4.53   0.000      1.33921    2.089209
       _rcs1 |   3.599906    .378516    12.18   0.000      2.92948    4.423761
       _rcs2 |   1.060463   .0826989     0.75   0.452     .9101551    1.235593
       _rcs3 |   1.003119    .025771     0.12   0.904     .9538591    1.054922
       _rcs4 |   1.024733   .0172455     1.45   0.147     .9914835    1.059097
    _rcs_tx1 |   .8110557   .0994424    -1.71   0.088     .6378023    1.031372
    _rcs_tx2 |   .9947941   .0843064    -0.06   0.951     .8425504    1.174547
       _cons |   .4795485   .0428071    -8.23   0.000     .4025777    .5712358
------------------------------------------------------------------------------
. test _rcs_tx1 _rcs_tx2

 ( 1)  [xb]_rcs_tx1 = 0
 ( 2)  [xb]_rcs_tx2 = 0

           chi2(  2) =    4.92
```

```
        Prob > chi2 =    0.0853
. predict hr, hrnumerator(tx 1) ci
. twoway (rarea hr_lci hr_uci _t if hr_uci<5, sort color(gs12)) (line hr _t if
> hr_uci<5, sort), legend(off) xtitle("Time since diagnosis") ytitle("Hazard ra
> tio") name(hr, replace) scheme(s2mono)
. graph export exam_2016_hr.eps, name(hr) replace
(file exam_2016_hr.eps written in EPS format)
. * the following line is only needed on Linux
. !! convert -density 300 exam_2016_hr.eps exam_2016_hr.png
```



We see that there is limited evidence for time-dependent hazards ($p = 0.09$ from the Wald test). We also see from the plot that the hazard ratio comparing chemotherapy+radiotherapy with conventional therapy is very high soon after diagnosis and then declines towards 1, with the confidence interval overlapping with 1 before the end of the first year.

Question 7

(a)

Advantages of using Poisson regression for Questions 5–6 include: (i) Poisson regression readily models for multiple time scales, where we could split on attained age and time since diagnosis and then model for main effects and interactions between those time scales and interactions between a time scale and other covariates; (ii) it is simpler to predict rates from Poisson regression, as the analysis is done on that scale.

Disadvantages of using Poisson regression include: (i) the need to split on the time scales, which may increase the size of the computational problem; (ii) the need to specify a functional form for the primary time scale using parametric functions, rather than using Cox regression's non-parametric formulation; (iii) crude time splitting will assume that rates are piece-wise constant, which may not be appropriate; (iv) risk calculations for Poisson regression require that the risk period involves constant rates or numerical integration.

(b)

Assuming that the follow-up time has been split for within one year of diagnosis and from one year of diagnosis, we can model the rate using:

$$\log(\lambda(t|\text{tx})) = \beta_0 + \beta_1 I(t < 1) + \beta_2 I(t \geq 1) + \beta_3 I(\text{tx} = 1) + \beta_4 I(\text{tx} = 1 \ \& \ t \geq 1)$$

A better formulation would be to include more time-splits for time since diagnosis. If we let time cuts be represented by t_j where $t_0 = 0$, then

$$\log(\lambda(t|\text{tx})) = \beta_0 + \sum_j \beta_j I(t_{j-1} < t \leq t_j) + \beta_{\text{tx}} I(\text{tx} = 1) + \beta_{\text{tx}:t} I(\text{tx} = 1 \ \& \ t \geq 1)$$

We could also model using splines. Any similar formulation was accepted, including different formulations for the time-dependent hazard ratios.