Biostat III Examination 2016 Answers

Mark Clements

May 17, 2016

Set-up

```
. global folder 2
. 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)      =     537.74
                                                  Prob > chi2     =     0.0000
Log likelihood = -887.51212                       Pseudo R2       =     0.2325

------------------------------------------------------------------------------
          lc |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |    2.16434   .2021694     8.27   0.000     1.802251    2.599175
         age |    1.09532   .0045813    21.77   0.000     1.086377    1.104336
       _cons |   1.76e-06   4.85e-07   -48.21   0.000     1.03e-06    3.02e-06
      ln(pt) |          1  (exposure)
```

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

Poisson regression                                Number of obs   =        360
                                                  LR chi2(2)      =    1348.60
                                                  Prob > chi2     =     0.0000
Log likelihood = -482.07822                       Pseudo R2       =     0.5831

------------------------------------------------------------------------------
          lc |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     smoking |   18.44911   2.225889    24.16   0.000      14.5639    23.37077
         age |   1.101785   .0046853    22.79   0.000      1.09264    1.111006
       _cons |   3.88e-07   1.14e-07   -50.44   0.000     2.19e-07    6.89e-07
      ln(pt) |          1  (exposure)
------------------------------------------------------------------------------
. poisson lc asbestos age, exposure(pt) nolog irr

Poisson regression                                Number of obs   =        360
                                                  LR chi2(2)      =     581.88
                                                  Prob > chi2     =     0.0000
Log likelihood = -865.43893                       Pseudo R2       =     0.2516

------------------------------------------------------------------------------
          lc |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    asbestos |   3.678602   .3897509    12.29   0.000     2.988803    4.527603
         age |    1.09465   .0045756    21.64   0.000     1.085719    1.103655
       _cons |   2.37e-06   6.34e-07   -48.49   0.000     1.41e-06    4.01e-06
      ln(pt) |          1  (exposure)
------------------------------------------------------------------------------
```

The age-adjusted incidence rate ratio for sex is 2.16 (95% confidence interval (CI): 1.80, 2.60). This association is highly significant ($p < 0.001$).

The age-adjusted incidence rate ratio for smoking is 18.45 (95% confidence interval (CI): 14.56, 23.37). This association is highly significant ($p < 0.001$).

The age-adjusted incidence rate ratio for asbestos is 3.68 (95% confidence interval (CI): 2.99, 4.53). 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)      =     547.57
                                                  Prob > chi2     =     0.0000
Log likelihood = -882.59664                       Pseudo R2       =     0.2368

------------------------------------------------------------------------------
          lc |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   2.150607   .2008611     8.20   0.000     1.790858    2.582624
             |
       ageQ5 |
          1  |          1  (base)
          2  |   4.315951   .8033874     7.86   0.000     2.996613    6.216163
          3  |    8.67328   1.573123    11.91   0.000     6.078495    12.37573
          4  |   19.63053   3.545159    16.48   0.000     13.77875    27.96752
```

```
           5  |   25.70656    5.75686    14.50   0.000     16.57377    39.87185
              |
       _cons  |   .0000606    .0000106   -55.73   0.000     .0000431    .0000853
      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)      =     555.20
                                                  Prob > chi2     =     0.0000
Log likelihood = -878.77969                       Pseudo R2       =     0.2401

------------------------------------------------------------------------------
          lc |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   2.156376   .2014116     8.23   0.000     1.795643    2.589579
  ageSpline1 |   1.162982   .0264366     6.64   0.000     1.112304    1.215968
  ageSpline2 |   .8930203     .05706    -1.77   0.077     .7879042     1.01216
  ageSpline3 |   1.200491   .2138047     1.03   0.305     .8467628    1.701986
       _cons |   8.71e-08   9.80e-08   -14.44   0.000     9.59e-09    7.91e-07
      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)      =    1478.37
                                                  Prob > chi2     =     0.0000
Log likelihood = -417.19618                       Pseudo R2       =     0.6392

------------------------------------------------------------------------------
          lc |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.104206   .0047395    23.09   0.000     1.094955    1.113534
         sex |   1.447719   .1369134     3.91   0.000     1.202775    1.742545
     smoking |    17.6259    2.13535    23.68   0.000     13.90048    22.34977
    asbestos |   3.419963   .3655378    11.50   0.000     2.773588    4.216974
       _cons |   2.37e-07   7.20e-08   -50.26   0.000     1.31e-07    4.30e-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 1.104 (95% CI: 1.095, 1.113) per year of age, indicating a rapid rise with increasing age. Males have higher rates of disease even after adjustment for other covariates (RR=1.45, 95% CI: 1.20, 1.74). Smoking is strongly associated with lung cancer incidence (RR=17.63, 95% CI: 13.90, 22.35). 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 |253.882356  20.345998 | 274.22835
           |     92.58       7.42 |    100.00
-----------+----------------------+----------
         1 | 78.888329  6.8833169 | 85.771646
           |     91.97       8.03 |    100.00
-----------+----------------------+----------
     Total | 332.77068  27.229315 |       360
           |     92.44       7.56 |    100.00
```

We see that the prevalence of exposure to asbestos is similar or slightly lower among never smokers (7.4%) and current smokers (8.0%). 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 the main effects model, we see that the rate ratio for asbestos changed from 3.68 to 3.42 (7% reduction), and the rate ratio for smoking changed from 18.45 to 17.63 (4% reduction). Again, there is limited evidence for confounding between smoking and asbestos.

Question 3

(a)

A regression model formula is

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

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)      =    1480.12
                                                  Prob > chi2     =     0.0000
Log likelihood = -416.31897                       Pseudo R2       =     0.6400
```

```
------------------------------------------------------------------------------
          lc |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    1.10411   .0047376    23.08   0.000     1.094863    1.113435
         sex |   1.443798   .1364852     3.89   0.000     1.199612     1.73769
   1.smoking |   19.29638   2.734882    20.88   0.000     14.61621    25.47516
  1.asbestos |   4.636403   1.143001     6.22   0.000     2.859806    7.516676
             |
     smoking#|
    asbestos |
        1 1  |   .6923611   .1887473    -1.35   0.177      .405773     1.18136
             |
       _cons |   2.21e-07   6.84e-08   -49.54   0.000     1.21e-07    4.06e-07
      ln(pt) |          1  (exposure)
------------------------------------------------------------------------------
. est store ModelB
. lrtest ModelA ModelB

Likelihood-ratio test                                 LR chi2(1)  =      1.75
(Assumption: ModelA nested in ModelB)                 Prob > chi2 =    0.1853
```

Comparing Model A with Model B, we see that there is little 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 0.18. Second, we see that the likelihood ratio test is also not significant, with $p = 0.19$.

(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. We can calculate the rate from the regression estimates, however we need to take account of the covariance terms to calculate the confidence interval, 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) |   .0091883    .000977   -44.11   0.000     .0074598    .0113173
------------------------------------------------------------------------------
```

This shows that the incidence rate is 9.19 (95% CI: 7.46, 10.32) per 1000 person-years.

We can do the same analysis using the `margins` command:

```
. margins smoking##asbestos, predict(ir) at(age=62 sex=1)

Predictive margins                              Number of obs     =        360
Model VCE    : OIM

Expression   : Predicted incidence rate, predict(ir)
at           : age             =          62
               sex             =           1

------------------------------------------------------------------------------
```

```
             |          Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     smoking |
          0  |    .000418   .0000735     5.69   0.000      .000274    .0005621
          1  |   .0060253   .0005059    11.91   0.000     .0050337    .0070169
             |
    asbestos |
          0  |   .0015053   .0000937    16.06   0.000     .0013216    .0016891
          1  |    .004938   .0004955     9.97   0.000     .0039668    .0059092
             |
     smoking#|
    asbestos |
        0 0  |   .0001483   .0000205     7.24   0.000     .0001082    .0001885
        0 1  |   .0006877   .0001446     4.75   0.000     .0004042    .0009712
        1 0  |   .0028623   .0001829    15.65   0.000     .0025039    .0032208
        1 1  |   .0091883    .000977     9.40   0.000     .0072734    .0111032
------------------------------------------------------------------------------
```

Finally, we could also do this analysis with the `predict` command.

Part 2

Question 4

We read in the data using the following:

```
. display "Folder = $folder"
Folder = 2
. import delimited "http://biostat3.net/download/exams/2016/$folder/survival.cs
> v", clear
(8 vars, 511 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 |     288    63.41218    .5391431    9.149561      62.351    64.47336
       1 |     223    62.78371     .667598    9.969363    61.46807    64.09935
---------+--------------------------------------------------------------------
combined |     511    63.13791     .420768     9.51159    62.31126    63.96457
---------+--------------------------------------------------------------------
    diff |            .6284714    .8488039               -1.039119    2.296062
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   0.7404
Ho: diff = 0                                     degrees of freedom =      509

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.7703         Pr(|T| > |t|) = 0.4594          Pr(T > t) = 0.2297
. ranksum age, by(tx)
```

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

          tx |      obs    rank sum    expected
-------------+---------------------------------
           0 |      288       74574       73728
           1 |      223       56242       57088
-------------+---------------------------------
    combined |      511      130816      130816

unadjusted variance  2740224.00
adjustment for ties        0.00
                     ----------
adjusted variance    2740224.00

Ho: age(tx==0) = age(tx==1)
             z =   0.511
    Prob > |z| =   0.6093
```

We find no evidence that age differs by treatment modality ($p = 0.46$ for the t-test and $p = 0.61$ for the Wilcoxon test). For the other variables:

```
. tab tx sex, chi row

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

           |          sex
        tx |         0          1 |     Total
-----------+----------------------+----------
         0 |       107        181 |       288
           |     37.15      62.85 |    100.00
-----------+----------------------+----------
         1 |        67        156 |       223
           |     30.04      69.96 |    100.00
-----------+----------------------+----------
     Total |       174        337 |       511
           |     34.05      65.95 |    100.00

          Pearson chi2(1) =   2.8277   Pr = 0.093
. tab tx smoking, chi row

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

           |        smoking
        tx |         0          1 |     Total
-----------+----------------------+----------
         0 |        41        247 |       288
           |     14.24      85.76 |    100.00
-----------+----------------------+----------
```

```
         1 |        41        182 |       223 
           |     18.39      81.61 |    100.00 
-----------+----------------------+----------
     Total |        82        429 |       511 
           |     16.05      83.95 |    100.00 

          Pearson chi2(1) =   1.6064   Pr = 0.205
. tab tx asbestos, chi row

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

           |       asbestos
        tx |         0          1 |     Total
-----------+----------------------+----------
         0 |       224         64 |       288 
           |     77.78      22.22 |    100.00 
-----------+----------------------+----------
         1 |       172         51 |       223 
           |     77.13      22.87 |    100.00 
-----------+----------------------+----------
     Total |       396        115 |       511 
           |     77.50      22.50 |    100.00 

          Pearson chi2(1) =   0.0302   Pr = 0.862
```

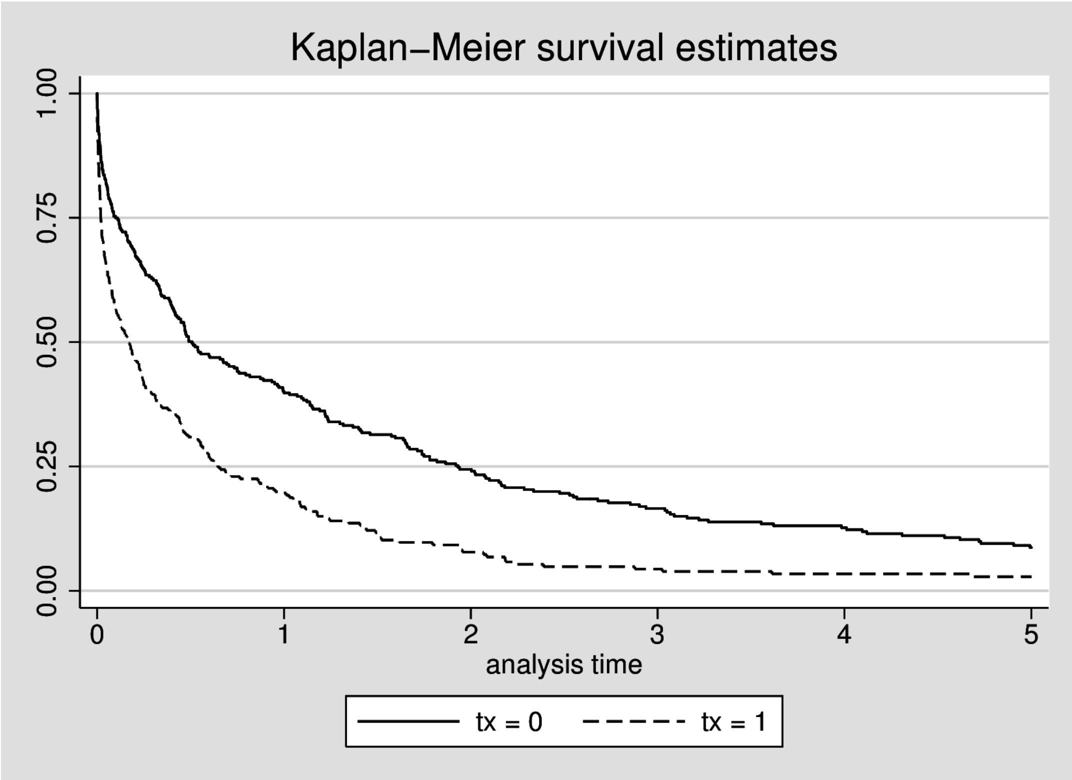
We find little evidence that randomisation varied by sex ($p = 0.09$), by smoking ($p = 0.21$) or by asbestos exposure ($p = 0.86$). We could check for potential confounding by sex in the survival analysis.

(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

------------------------------------------------------------------------------
      511  total observations
        0  exclusions
------------------------------------------------------------------------------
      511  observations remaining, representing
      511  subjects
      468  failures in single-failure-per-subject data
  486.346  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_$folder.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     |       255         316.29
1     |       213         151.71
------+-------------------------
Total |       468         468.00

            chi2(1) =      37.72
            Pr>chi2 =     0.0000
. 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       112      172                  0.3980    0.0290      0.3410    0.4543
       2        67       42                  0.2444    0.0258      0.1956    0.2961
       3        44       21                  0.1655    0.0225      0.1242    0.2120
       4        33       10                  0.1269    0.0203      0.0905    0.1697
       5        21       10                  0.0870    0.0174      0.0569    0.1250
tx=1
       1        43      178                  0.1970    0.0269      0.1472    0.2521
       2        17       25                  0.0777    0.0184      0.0467    0.1188
       3        10        7                  0.0437    0.0141      0.0216    0.0776
       4         8        2                  0.0340    0.0126      0.0152    0.0652
       5         5        1                  0.0283    0.0117      0.0114    0.0584
-------------------------------------------------------------------------------
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 9% (95% CI: 6, 13) survived for those on conventional treatment and 3% (95% CI: 1, 6) 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 =          511                     Number of obs   =       511
No. of failures =          468
Time at risk    =  486.3459971
                                                   LR chi2(3)      =     37.25
Log likelihood  =   -2518.2339                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          tx |   1.763393   .1680097     5.95   0.000     1.463019    2.125439
         sex |   1.031313   .1011442     0.31   0.753     .8509622    1.249886
         age |   .9946115   .0046482    -1.16   0.248     .9855428    1.003764
------------------------------------------------------------------------------
```

```
. stcox tx sex, nolog

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

Cox regression -- no ties

No. of subjects =          511                     Number of obs   =       511
No. of failures =          468
Time at risk    =  486.3459971
                                                   LR chi2(2)      =     35.92
Log likelihood  =   -2518.9007                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          tx |   1.766735   .1683261     5.97   0.000     1.465794    2.129462
         sex |   1.032023   .1012402     0.32   0.748     .8515052    1.250809
------------------------------------------------------------------------------
. stcox tx age, nolog

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

Cox regression -- no ties

No. of subjects =          511                     Number of obs   =       511
No. of failures =          468
Time at risk    =  486.3459971
                                                   LR chi2(2)      =     37.15
Log likelihood  =   -2518.2835                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          tx |   1.770119   .1673069     6.04   0.000     1.470785    2.130373
         age |   .9945999   .0046504    -1.16   0.247      .985527    1.003756
------------------------------------------------------------------------------
. stcox tx, nolog

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

Cox regression -- no ties

No. of subjects =          511                     Number of obs   =       511
No. of failures =          468
Time at risk    =  486.3459971
                                                   LR chi2(1)      =     35.81
Log likelihood  =   -2518.9524                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          tx |   1.773758    .167595     6.07   0.000       1.4739    2.134621
```

```
------------------------------------------------------------------------------
```

Adjusting for treatment modality, there is no evidence that either sex or age are associated with survival, with Wald test p-values of 0.75 and 0.25 for sex and age, respectively. 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 1.77 (95% CI: 1.47, 2.13), suggesting that the average hazard ratio for chemotherapy+radiotherapy is high over the five-year period.

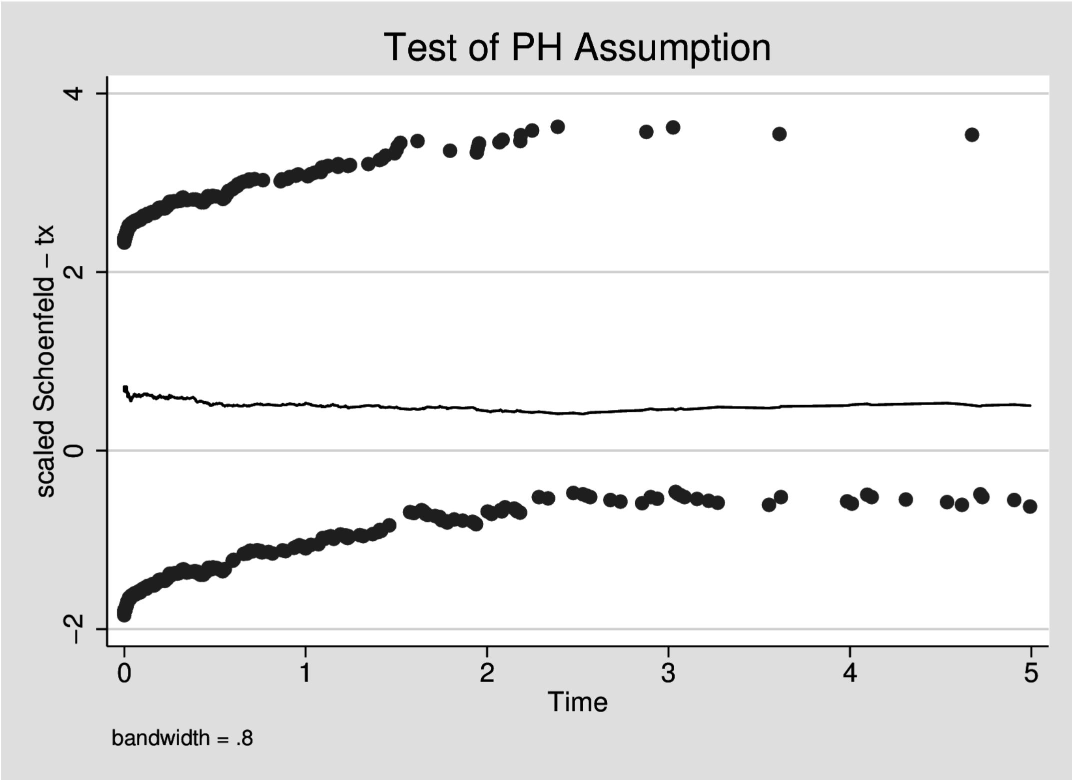
For the time scale, we have initially used time since cancer diagnosis. There is a strong association between time since diagnosis and survival, suggesting that this is the best choice of primary time scale. Moreover, there is a suggestion of non-proportional hazards, with a higher rate ratio in the 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)

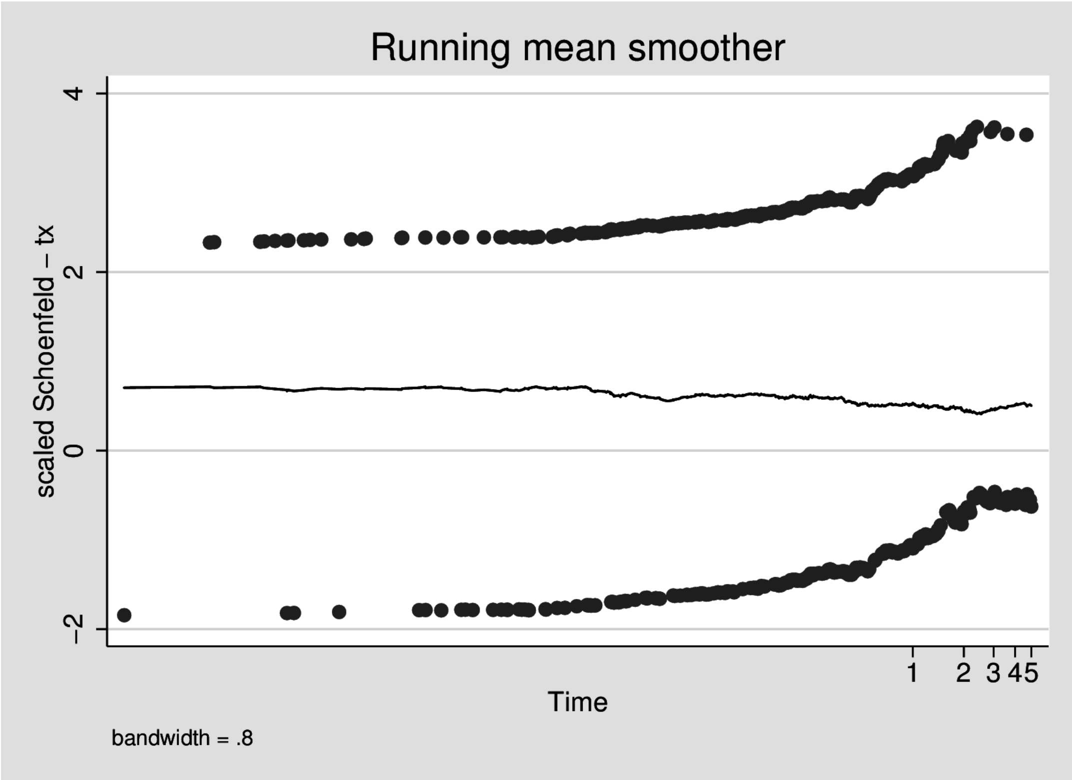
For an analysis of scaled Schoenfeld residuals, we use:

```
. estat phtest, detail

      Test of proportional-hazards assumption

      Time:  Time
      ----------------------------------------------------------------
                  |       rho            chi2       df       Prob>chi2
      ------------+---------------------------------------------------
      tx          |     -0.06834          2.13        1         0.1448
      ------------+---------------------------------------------------
      global test |                       2.13        1         0.1448
      ----------------------------------------------------------------
. 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_$folder.png
```



This shows that there is little evidence ($p = 0.14$) that the hazard ratio decreases with increasing time since diagnosis: the scaled residuals and linear time have a correlation of -0.07. 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          |     -0.06680         2.03        1         0.1541
      ------------+---------------------------------------------------
      global test |                      2.03        1         0.1541
      ----------------------------------------------------------------
. 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_$folder.p
> ng
```



Again, there is little evidence for non-proportionality ($p = 0.15$).

(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)
(153 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 =          511                     Number of obs   =        664
No. of failures =          468
Time at risk    =  486.3459971
                                                   LR chi2(3)      =      36.24
Log likelihood  =   -2518.7386                     Prob > chi2     =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   1.032486    .101294     0.33   0.745     .8518739    1.251391
        1.tx |   1.818564   .1965733     5.53   0.000     1.471363    2.247694
  1.timeband |   7.525831          .        .       .            .           .
             |
```

```
 tx#timeband |
        1 1  |   .8782778   .2012448    -0.57   0.571     .5605201    1.376171
------------------------------------------------------------------------------
. stcox tx sex c.tx#c.timeband, nolog

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

Cox regression -- no ties

No. of subjects =          511                     Number of obs   =       664
No. of failures =          468
Time at risk    =  486.3459971
                                                   LR chi2(3)      =     36.24
Log likelihood  =   -2518.7386                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          tx |   1.818564   .1965733     5.53   0.000     1.471363    2.247694
         sex |   1.032486    .101294     0.33   0.745     .8518739    1.251391
             |
        c.tx#|
  c.timeband |   .8782778   .2012448    -0.57   0.571     .5605201    1.376171
------------------------------------------------------------------------------
. stcox c.tx#i.timeband, nolog

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

Cox regression -- no ties

No. of subjects =          511                     Number of obs   =       664
No. of failures =          468
Time at risk    =  486.3459971
                                                   LR chi2(2)      =     36.13
Log likelihood  =   -2518.7918                     Prob > chi2     =    0.0000

-------------------------------------------------------------------------------
           _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+----------------------------------------------------------------
timeband#c.tx |
          0   |   1.825656    .196156     5.60   0.000      1.47898    2.253593
          1   |   1.604385   .3246888     2.34   0.019     1.079061    2.385455
-------------------------------------------------------------------------------
```

This model provides little or no evidence that the hazard ratio is time-dependent ($p = 0.57$). The hazard ratio in the first year is 1.83 (95% CI: 1.48, 2.25), while the hazard ratio after the first year is 1.60 (95% CI: 1.08, 2.39).

(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 =          511                     Number of obs   =       664
No. of failures =          468
Time at risk    =  486.3459971
                                                   LR chi2(2)      =     36.13
Log likelihood  =   -2518.7918                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
main         |
          tx |   1.825656    .196156     5.60   0.000      1.47898    2.253593
-------------+----------------------------------------------------------------
tvc          |
          tx |   .8787997   .2013587    -0.56   0.573     .5608603    1.376972
------------------------------------------------------------------------------
Note: variables in tvc equation interacted with _t>=1
```

Again, we find little evidence for a time-dependent hazard ratio ($p = 0.57$). We can model for a time-dependent hazard ratio that depends on time:

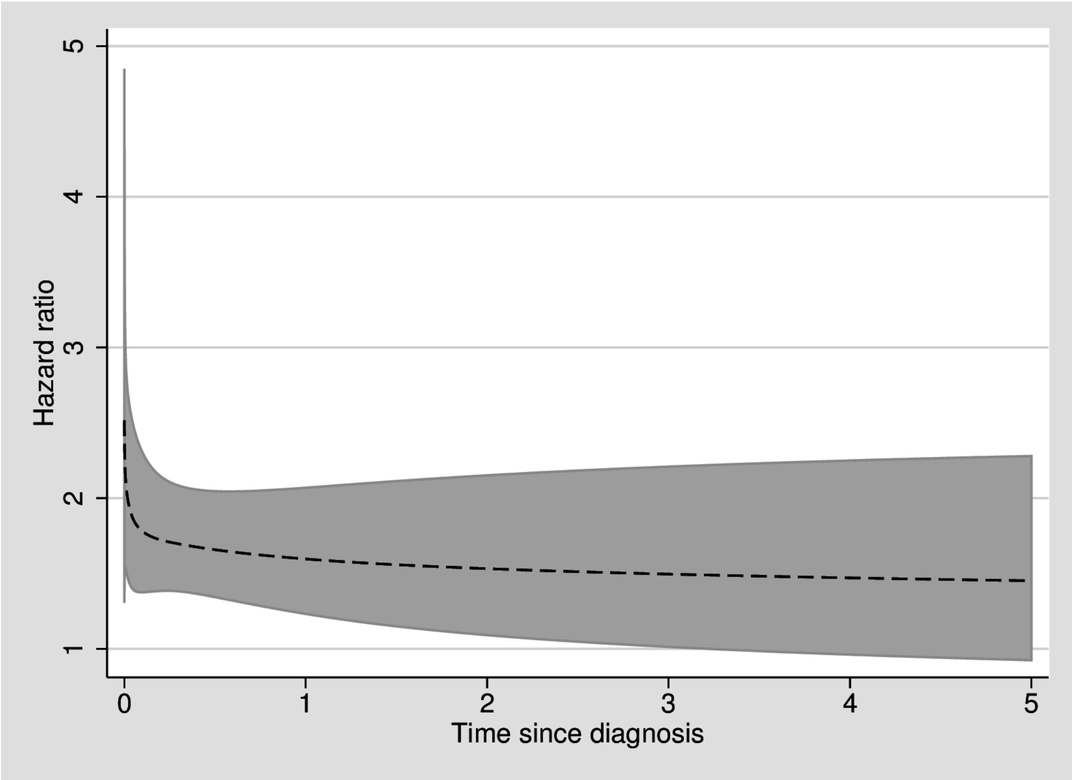
```
. stcox tx, nolog tvc(tx) texp(_t)

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


Cox regression -- no ties

No. of subjects =          511                     Number of obs   =       664
No. of failures =          468
Time at risk    =  486.3459971
                                                   LR chi2(2)      =     38.39
Log likelihood  =   -2517.6658                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
main         |
          tx |   1.972851   .2293675     5.84   0.000      1.57084    2.477745
-------------+----------------------------------------------------------------
tvc          |
          tx |   .8410601   .0942332    -1.54   0.122     .6752391    1.047602
------------------------------------------------------------------------------
Note: variables in tvc equation interacted with _t
```

The interpretation of this model is as follows: the hazard ratio at time 0 is 1.97 (95% CI: 1.57, 2.48); for each year since diagnosis, the rate tends to decrease by 1-0.84=16% (RR=0.84, 95% CI: 0.68, 1.05), although this trend is not significant ($p = 0.12$, as per the Schoenfeld test).

(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 = -1106.6321                     Number of obs   =        664

------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
          tx |   1.903862   .2047045     5.99   0.000     1.542105    2.350484
       _rcs1 |   3.441882   .2972226    14.31   0.000     2.905967    4.076629
       _rcs2 |    1.06212   .0650871     0.98   0.325     .9419153    1.197666
       _rcs3 |   .9446278   .0229253    -2.35   0.019     .9007469    .9906464
       _rcs4 |   1.009233   .0168565     0.55   0.582     .9767296    1.042818
    _rcs_tx1 |   .8817852   .0949035    -1.17   0.242     .7140867    1.088867
    _rcs_tx2 |   1.001463   .0724086     0.02   0.984     .8691419     1.15393
       _cons |   .4874301   .0388629    -9.01   0.000     .4169134    .5698739
------------------------------------------------------------------------------
. test _rcs_tx1 _rcs_tx2

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

           chi2(  2) =    2.02
         Prob > chi2 =    0.3651
. 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_$folder.png
```



We see that there is limited evidence for time-dependent hazards ($p = 0.37$ from the Wald test). We also see from the plot that the hazard ratio looks comparatively stable across the follow-up period.

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.