Biostat III Examination 2017 Answers

Mark Clements

April 10, 2017

Part 1

Some of the questions provided latitude in the analytical approach (e.g. 3(b)) and some of the questions required interpretation (e.g. defining 'safe' in 5(c)). As a general comment, it was important to provide commentary on the results, where code and output were not sufficient to get full marks. Moreover, it was important to define the notation used in the equations.

Initially, we (i) set the line size, (ii) change to the data folder, (iii) read in the dataset, (iv) create categorical variables for PSA and age at study entry, and (v) save a copy of the file for `use` later.

```
. set linesize 80
. cd /home/marcle/repos/biostat3_2014/exam/2017
/home/marcle/repos/biostat3_2014/exam/2017
. import delimited "psa.csv", clear
(7 vars, 100,000 obs)
. egen psa_cat = cut(psa), at(0,1,2,3,10,17586) label
. egen age_cat = cut(start_age), at(50,60,70,80) label
. saveold psa, version(11) replace
(saving in Stata 12 format, which Stata 11 can read)
file psa.dta saved
```

Question 1

If we consider the lower PSA categories, then 45.4% (95% CI: 45.0, 45.8) of those aged 50-59 years at study entry have a PSA value less than 1 ng/mL; in contrast, 32.1% (95% CI: 31.6, 32.6) of those aged 60-69 years and 22.3% (95% CI: 21.7, 23.0) of those aged 70-79 years have PSA $<$ 1 ng/mL. Based on the chi-square test, we find strong evidence for differences in the PSA categories by age categories, although the small p-value may be also an indication of the large cell sizes. Note that the confidence intervals for the proportions were not expected, but they would be useful for a description of the sample for a scientific article.

```
. use psa, clear
. tab age_cat psa_cat, row chi

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

           |                         psa_cat
   age_cat |        0-         1-         2-         3-        10- |     Total
-----------+-------------------------------------------------------+----------
       50- |    23,233     13,613      5,750      7,051      1,526 |    51,173
           |     45.40      26.60      11.24      13.78       2.98 |    100.00
-----------+-------------------------------------------------------+----------
       60- |    10,072      7,666      4,173      7,386      2,096 |    31,393
```

```
           |     32.08      24.42      13.29      23.53       6.68 |    100.00
-----------+-------------------------------------------------------+----------
       70- |     3,895      3,389      2,292      5,379      2,479 |    17,434
           |     22.34      19.44      13.15      30.85      14.22 |    100.00
-----------+-------------------------------------------------------+----------
     Total |    37,200     24,668     12,215     19,816      6,101 |   100,000
           |     37.20      24.67      12.21      19.82       6.10 |    100.00

          Pearson chi2(8) =  7.5e+03   Pr = 0.000
. quietly capture tab psa_cat, gen(psa_cat)
. bysort age_cat: ci proportions psa_cat1

--------------------------------------------------------------------------------
-> age_cat = 50-

                                                         -- Binomial Exact --
    Variable |        Obs  Proportion    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
    psa_cat1 |     51,173     .454009    .0022009        .4496882     .458335

--------------------------------------------------------------------------------
-> age_cat = 60-

                                                         -- Binomial Exact --
    Variable |        Obs  Proportion    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
    psa_cat1 |     31,393    .3208359    .0026346         .315673    .3260319

--------------------------------------------------------------------------------
-> age_cat = 70-

                                                         -- Binomial Exact --
    Variable |        Obs  Proportion    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
    psa_cat1 |     17,434     .223414    .0031547        .2172489    .2296716
```

For the formal test for the association, we could also use a variety of other methods, including: non-parametric tests on the PSA values; linear regression or analysis of variance for the log(PSA) values; and binomial regression for a cut-point in PSA values. Some of these results are shown below. Note that the long tail in the PSA values suggests using log(PSA) values; the assumption here is that the measurement error is also on the log-scale. For these alternative approaches, they all provide strong evidence that age at the initial PSA test is strongly associated with the PSA values. From the linear regression, we estimate that a one year increase in age will lead to 4% increase in the PSA value.

```
.
. kwallis psa, by(age_cat)
at least two populations are required
r(498);
.
. capture drop ln_psa
. capture drop start_age_50
. gen start_age_50 = start_age - 50
. gen ln_psa = ln(psa)
. reg ln_psa i.age_cat, base
note: 1.age_cat omitted because of collinearity

      Source |       SS           df       MS      Number of obs   =    31,393
-------------+----------------------------------   F(0, 31392)     =      0.00
       Model |           0         0           .   Prob > F        =         .
```

```
    Residual |  47741.7414    31,392  1.5208251   R-squared       =    0.0000
-------------+----------------------------------   Adj R-squared   =    0.0000
       Total |  47741.7414    31,392  1.5208251   Root MSE        =    1.2332

------------------------------------------------------------------------------
      ln_psa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     age_cat |
        60-  |          0  (omitted)
             |
       _cons |   .5682677   .0069602    81.65   0.000     .5546254      .58191
------------------------------------------------------------------------------
. display "Proportional change in PSA values compared with 50-59 years: " exp(.
> 3688477) " and " exp(.7930282) " for those aged 60-69 years and 70-79 years,
> respectively."
Proportional change in PSA values compared with 50-59 years: 1.4460674 and 2.210
> 0789 for those aged 60-69 years and 70-79 years, respectively.
. reg ln_psa start_age_50, base

      Source |       SS           df       MS      Number of obs   =    31,393
-------------+----------------------------------   F(1, 31391)     =    280.10
       Model |  422.232067         1  422.232067   Prob > F        =    0.0000
    Residual |  47319.5093    31,391  1.50742281   R-squared       =    0.0088
-------------+----------------------------------   Adj R-squared   =    0.0088
       Total |  47741.7414    31,392  1.5208251   Root MSE        =    1.2278

------------------------------------------------------------------------------
      ln_psa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
start_age_50 |   .0407087   .0024324    16.74   0.000     .0359412    .0454763
       _cons |   .0023887   .0345144     0.07   0.945    -.0652608    .0700382
------------------------------------------------------------------------------
. display "Proportional change in PSA per unit change in start age: " exp(.0407
> 052)
Proportional change in PSA per unit change in start age: 1.041545
.
. capture drop psa_ge_10
. gen psa_ge_10 = (psa>=10)
. logit psa_ge_10 i.age_cat, nolog base or
note: 1.age_cat omitted because of collinearity

Logistic regression                             Number of obs     =     31,393
                                                LR chi2(0)        =       0.00
                                                Prob > chi2       =          .
Log likelihood = -7697.3549                     Pseudo R2         =     0.0000

------------------------------------------------------------------------------
   psa_ge_10 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     age_cat |
        60-  |          1  (omitted)
             |
       _cons |   .0715432   .0016176  -116.65   0.000     .0684419    .0747849
------------------------------------------------------------------------------
```

Question 2

We then restrict the dataset and `stset` the dataset for time since study entry. We note that one individual has an age of prostate cancer diagnosis and age of death that precedes the age of study entry; that individual should be excluded from the analyses in Parts 1 and 2.

```
. use psa, clear
. keep if start_age>=50 & start_age<70 & psa<3
(35,493 observations deleted)
. stset age_dx, fail(event_dx==1) origin(start_age)

     failure event:  event_dx == 1
obs. time interval:  (origin, age_dx]
 exit on or before:  failure
    t for analysis:  (time-origin)
            origin:  time start_age

------------------------------------------------------------------------------
      64507  total observations
          1  observation ends on or before enter()
------------------------------------------------------------------------------
      64506  observations remaining, representing
       2908  failures in single-record/single-failure data
  713493.27  total analysis time at risk and under observation
                                                at risk from t =         0
                                     earliest observed entry t =         0
                                          last observed exit t =  12.65369
. list if age_dx < start_age

       +------------------------------------------------------------+
48357. |    id | start_~e |    age_dx | event_dx |   age_dth | event_~h |
       | 76042 |       55 | 21.98011 |        0 | 21.98011 |        2 |
       |---------------------------------------------------+----------|
       |      psa | psa_cat | age_cat | _st | _d | _origin | _t | _t0 |
       | .5194139 |      0- |     50- |   0 |  . |      55 |  . |   . |
       +------------------------------------------------------------+
. drop if age_dx < start_age
(1 observation deleted)
```

(a)

Using Poisson regression to model the rate of prostate cancer incidence by age and PSA categories, we can use `streg`, `poisson` or `glm` commands. All three approaches should give the same estimates. For men with PSA below 3 ng/mL, there was some evidence that men aged 60-69 years had slightly higher incidence rates than men aged 50-59 years (incidence rate ratio (IRR) = 1.08, 95% CI: 1.00, 1.16, p=0.06). There was much stronger evidence that the incidence rates rose with increasing PSA categories: compared with men whose initial PSA value was less than 1 ng/mL, men with values between 1 and 2 ng/mL had an IRR of 2.79 (95% CI: 2.54, 3.08; p<0.001), and men with PSA values between 2 and 3 ng/mL had 6.09 times the incidence rate (95% CI: 5.52, 6.71; p<0.001).

```
. use psa, clear
. keep if start_age>=50 & start_age<70 & psa<3
(35,493 observations deleted)
. stset age_dx, fail(event_dx==1) origin(start_age)

     failure event:  event_dx == 1
obs. time interval:  (origin, age_dx]
 exit on or before:  failure
    t for analysis:  (time-origin)
```

```
           origin:  time start_age

------------------------------------------------------------------------------
      64507  total observations
          1  observation ends on or before enter()
------------------------------------------------------------------------------
      64506  observations remaining, representing
       2908  failures in single-record/single-failure data
  713493.27  total analysis time at risk and under observation
                                                at risk from t =         0
                                     earliest observed entry t =         0
                                          last observed exit t =  12.65369
.
. streg i.age_cat i.psa_cat, dist(exp) base nolog

         failure _d:  event_dx == 1
   analysis time _t:  (age_dx-origin)
             origin:  time start_age

Exponential regression -- log relative-hazard form

No. of subjects =       64,506                  Number of obs    =     64,506
No. of failures =        2,908
Time at risk    =  713493.2702
                                                LR chi2(3)       =    1447.12
Log likelihood  =   -12646.774                  Prob > chi2      =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     age_cat |
        50-  |          1  (base)
        60-  |   1.076608   .0413881     1.92   0.055     .9984693    1.160861
             |
     psa_cat |
         0-  |          1  (base)
         1-  |   2.794569   .1385949    20.72   0.000     2.535713    3.079851
         2-  |   6.089436   .3030691    36.30   0.000     5.523484    6.713378
             |
       _cons |   .0016795   .0000695  -154.37   0.000     .0015487    .0018214
------------------------------------------------------------------------------
. testparm i(1 2).psa_cat

 ( 1)  [_t]1.psa_cat = 0
 ( 2)  [_t]2.psa_cat = 0

           chi2(  2) = 1329.04
         Prob > chi2 =    0.0000
.
. poisson _d i.age_cat i.psa_cat if _st==1, exposure(_t) nolog irr base

Poisson regression                              Number of obs     =     64,506
                                                LR chi2(3)        =    1447.12
                                                Prob > chi2       =     0.0000
Log likelihood = -12646.774                     Pseudo R2         =     0.0541

------------------------------------------------------------------------------
          _d |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
     age_cat |
        50-  |          1  (base)
        60-  |   1.076608   .0413881     1.92   0.055     .9984693    1.160861
             |
     psa_cat |
         0-  |          1  (base)
         1-  |   2.794573   .1385951    20.72   0.000     2.535716    3.079855
         2-  |   6.089444   .3030697    36.30   0.000      5.52349    6.713386
             |
       _cons |   .0016795   .0000695  -154.37   0.000     .0015487    .0018214
      ln(_t) |          1  (exposure)
------------------------------------------------------------------------------
. capture drop ln_pt
. gen ln_pt = ln(_t) if _st==1
(1 missing value generated)
. glm _d i.age_cat i.psa_cat if _st==1, family(poisson) offset(ln_pt) nolog efo
> rm base

Generalized linear models                         No. of obs      =     64,506
Optimization     : ML                             Residual df     =     64,502
                                                  Scale parameter =          1
Deviance         =  19477.54726                   (1/df) Deviance =   .3019681
Pearson          =  212182.7134                   (1/df) Pearson  =   3.289552

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   .3922356
Log likelihood   = -12646.77363                   BIC             =  -694850.7

------------------------------------------------------------------------------
             |                 OIM
          _d |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     age_cat |
        50-  |          1  (base)
        60-  |   1.076608   .0413881     1.92   0.055     .9984693    1.160861
             |
     psa_cat |
         0-  |          1  (base)
         1-  |   2.794573   .1385951    20.72   0.000     2.535716    3.079855
         2-  |   6.089444   .3030696    36.30   0.000      5.52349    6.713386
             |
       _cons |   .0016795   .0000695  -154.37   0.000     .0015487    .0018214
       ln_pt |          1  (offset)
------------------------------------------------------------------------------
```

We could also have used `poisson` or `glm` without using `stset`. This was a common cause of errors, either due to not including person-time in the analysis (such that the analysis was for counts and not for rates), or using the wrong person-time (e.g. using the age at diagnosis as the person-time). The individual with their diagnosis preceding their initial PSA value could cause problems here and should be excluded. The output is the same as before.

```
. use psa, clear
. keep if start_age>=50 & start_age<70 & psa<3
(35,493 observations deleted)
. drop if age_dx < start_age
(1 observation deleted)
```

```
. capture drop person_time
. gen person_time = age_dx - start_age
.
. poisson event_dx i.age_cat i.psa_cat, exposure(person_time) nolog irr base

Poisson regression                              Number of obs     =     64,506
                                                LR chi2(3)        =    1447.12
                                                Prob > chi2       =     0.0000
Log likelihood = -12646.774                     Pseudo R2         =     0.0541

------------------------------------------------------------------------------
    event_dx |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     age_cat |
        50-  |          1  (base)
        60-  |   1.076608   .0413881     1.92   0.055     .9984693    1.160861
             |
     psa_cat |
         0-  |          1  (base)
         1-  |   2.794573   .1385951    20.72   0.000     2.535716    3.079855
         2-  |   6.089444   .3030697    36.30   0.000      5.52349    6.713386
             |
       _cons |   .0016795   .0000695  -154.37   0.000     .0015487    .0018214
ln(person~e) |          1  (exposure)
------------------------------------------------------------------------------
. capture drop ln_pt
. gen ln_pt = ln(person_time)
. glm event_dx i.age_cat i.psa_cat, family(poisson) offset(ln_pt) nolog eform b
> ase

Generalized linear models                         No. of obs      =     64,506
Optimization     : ML                             Residual df     =     64,502
                                                  Scale parameter =          1
Deviance         =  19477.54726                   (1/df) Deviance =   .3019681
Pearson          =  212182.7134                   (1/df) Pearson  =   3.289552

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   .3922356
Log likelihood   = -12646.77363                   BIC             =  -694850.7

------------------------------------------------------------------------------
             |                 OIM
    event_dx |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     age_cat |
        50-  |          1  (base)
        60-  |   1.076608   .0413881     1.92   0.055     .9984693    1.160861
             |
     psa_cat |
         0-  |          1  (base)
         1-  |   2.794573   .1385951    20.72   0.000     2.535716    3.079855
         2-  |   6.089444   .3030696    36.30   0.000      5.52349    6.713386
             |
       _cons |   .0016795   .0000695  -154.37   0.000     .0015487    .0018214
       ln_pt |          1  (offset)
------------------------------------------------------------------------------
```

(b)

To assess the interactions, we fit three models. First, we fit a main effects model and store the estimates. We use `quietly` because the printed output is not used here. Second, we fit an interaction model to assess the size of the interactions. We then compare the first and second models for a formal test for interaction. Similarly, we use a Wald test to test for an interaction. Third, we re-parameterise the effects so that we can more easily describe the interactions. From these models, we find that there is strong evidence for an interaction, although the likelihood-ratio and Wald p-values are difficult to interpret due to the large cell sizes. There is clear evidence that the differences between PSA categories vary by age categories: for men aged 50-59 years, the incidence rate ratios for 1-2 and 2-3 ng/mL compared with 0-1 ng/mL are 3.48 (95% CI: 3.07, 3.94) and 7.79 (95% CI: 6.86, 8.83), respectively; in contrast, for men aged 60-69 years, the same IRRs were 1.89 (95% CI: 1.62, 2.21) and 4.03 (95% CI: 3.46, 4.69), respectively.

```
. use psa, clear
. keep if start_age>=50 & start_age<70 & psa<3
(35,493 observations deleted)
. quietly stset age_dx, fail(event_dx==1) origin(start_age)
. quietly streg i.age_cat i.psa_cat, dist(exp) base nolog
. quietly est store base
. streg i.age_cat##i.psa_cat, dist(exp) base nolog

         failure _d:  event_dx == 1
   analysis time _t:  (age_dx-origin)
             origin:  time start_age

Exponential regression -- log relative-hazard form

No. of subjects =       64,506                  Number of obs    =     64,506
No. of failures =        2,908
Time at risk    =  713493.2702
                                                LR chi2(5)       =    1494.40
Log likelihood  =   -12623.134                  Prob > chi2      =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     age_cat |
        50-  |          1  (base)
        60-  |   1.772385   .1424837     7.12   0.000     1.514011    2.074851
             |
     psa_cat |
         0-  |          1  (base)
         1-  |   3.480924   .2215444    19.60   0.000     3.072696    3.943388
         2-  |   7.785864   .5020712    31.83   0.000     6.861469    8.834796
             |
     age_cat#|
     psa_cat |
     60-#1-  |   .5435565   .0553788    -5.98   0.000     .4451663    .6636929
     60-#2-  |   .5175907   .0521871    -6.53   0.000     .4247784    .6306821
             |
       _cons |   .0014023   .0000731  -126.03   0.000     .0012661    .0015531
------------------------------------------------------------------------------
. lrtest base

Likelihood-ratio test                                 LR chi2(2)  =     47.28
(Assumption: base nested in .)                        Prob > chi2 =    0.0000
. testparm i1.age_cat#i(1 2).psa_cat
```

```
 ( 1)  [_t]1.age_cat#1.psa_cat = 0
 ( 2)  [_t]1.age_cat#2.psa_cat = 0

           chi2(  2) =   48.48
         Prob > chi2 =    0.0000
. streg i.age_cat i.age_cat#i.psa_cat, dist(exp) base nolog

         failure _d:  event_dx == 1
   analysis time _t:  (age_dx-origin)
             origin:  time start_age

Exponential regression -- log relative-hazard form

No. of subjects =       64,506                  Number of obs    =      64,506
No. of failures =        2,908
Time at risk    =  713493.2702
                                                LR chi2(5)       =     1494.40
Log likelihood  =   -12623.134                  Prob > chi2      =      0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     age_cat |
        50-  |          1  (base)
        60-  |   1.772385   .1424837     7.12   0.000     1.514011    2.074851
             |
     age_cat#|
     psa_cat |
     50-#1-  |   3.480924   .2215444    19.60   0.000     3.072696    3.943388
     50-#2-  |   7.785864   .5020712    31.83   0.000     6.861469    8.834796
     60-#1-  |   1.892079   .1505279     8.02   0.000     1.618901    2.211354
     60-#2-  |   4.029891   .3123558    17.98   0.000     3.461918    4.691046
             |
       _cons |   .0014023   .0000731  -126.03   0.000     .0012661    .0015531
------------------------------------------------------------------------------
```

(c)

For the first interaction model above, the regression equation is:

$$\begin{aligned}\log(\lambda(t|psa\_cat, age\_cat)) = \beta_0 + \beta_1 I(age\_cat = "60-")+ \\ \beta_2 I(psa\_cat = "1-")+ \\ \beta_3 I(psa\_cat = "2-")+ \\ \beta_4 I(psa\_cat = "1-" \,\&\, age\_cat = "60-")+ \\ \beta_5 I(psa\_cat = "2-" \,\&\, age\_cat = "60-")\end{aligned}$$

where $\lambda(t|psa\_cat, age\_cat)$ is the prostate cancer incidence rate at time t given psa_cat and age_cat, β_k are the regression parameters for $k = 0, \ldots, 5$, and $I()$ are indicator functions.

For the second interaction model above, the regression equation is:

$$\begin{aligned}\log(\lambda(t|psa\_cat, age\_cat)) = \beta_0 + \beta_1 I(age\_cat = "60-")+ \\ \beta_2 I(psa\_cat = "1-" \,\&\, age\_cat = "50-")+ \\ \beta_3 I(psa\_cat = "2-" \,\&\, age\_cat = "50-")+ \\ \beta_4 I(psa\_cat = "1-" \,\&\, age\_cat = "60-")+ \\ \beta_5 I(psa\_cat = "2-" \,\&\, age\_cat = "60-")\end{aligned}$$

(d)

Although not asked for, we can use the regression equation in (c) to define the rate equation for a man aged 62 years with a PSA value of 1.1 ng/mL: $\exp(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_4)$. We can calculate the predicted rate by hand from the fitted model. We can also obtain a 95% confidence interval using the `lincom` command. We find that the predicted incidence rate is 4.70 per 1000 person-years (95% CI: 4.26, 5.20).

```
. streg i.age_cat##i.psa_cat, dist(exp) base nolog nohr

         failure _d:  event_dx == 1
   analysis time _t:  (age_dx-origin)
             origin:  time start_age

Exponential regression -- log relative-hazard form

No. of subjects =       64,506                  Number of obs    =      64,506
No. of failures =        2,908
Time at risk    =  713493.2702
                                                LR chi2(5)       =     1494.40
Log likelihood  =   -12623.134                  Prob > chi2      =      0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     age_cat |
        50-  |          0  (base)
        60-  |    .572326     .080391     7.12   0.000     .4147626    .7298894
             |
     psa_cat |
         0-  |          0  (base)
         1-  |   1.247298    .0636453    19.60   0.000     1.122555     1.37204
         2-  |    2.05231     .064485    31.83   0.000     1.925922    2.178698
             |
    age_cat#|
     psa_cat |
     60-#1-  |  -.6096216    .1018824    -5.98   0.000    -.8093074   -.4099358
     60-#2-  |  -.6585705    .1008269    -6.53   0.000    -.8561877   -.4609534
             |
       _cons |  -6.569647    .0521286  -126.03   0.000    -6.671817   -6.467477
------------------------------------------------------------------------------
. display exp(-6.569647+.572326+1.247298+-.6096216)
.00470258
. lincom _cons+i1.age_cat+i1.psa_cat+i1.age_cat#1.psa_cat, eform

 ( 1)  [_t]1.age_cat + [_t]1.psa_cat + [_t]1.age_cat#1.psa_cat + [_t]_cons = 0

------------------------------------------------------------------------------
          _t |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .0047026     .000239  -105.44   0.000     .0042566    .0051952
------------------------------------------------------------------------------
```

(e)

The formula for the risk calculation is $1 - \exp(-\hat{\lambda}t)$. We can use the confidence interval for the rate $\hat{\lambda}$ with $t = 10$. The code is quite simple:

```
. display 1-exp(-10*.0047026)
.04593741
```

```
. display 1-exp(-10*.0042566)
.04167279
. display 1-exp(-10*.0051952)
.05062556
```

To do this in code, it is simpler to not use `eform` option, as `lincom` only returns the estimate and standard error, rather than the confidence interval. We re-run the `lincom` command and then use the returned values to calculate the confidence interval:

```
. lincom _cons+i1.age_cat+i1.psa_cat+i1.age_cat#1.psa_cat

 ( 1)  [_t]1.age_cat + [_t]1.psa_cat + [_t]1.age_cat#1.psa_cat + [_t]_cons = 0

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -5.359645   .0508329  -105.44   0.000    -5.459276   -5.260014
------------------------------------------------------------------------------
. display "Ten-year risk: " 1-exp(-exp(r(estimate))*10)
Ten-year risk: .04593717
. display "Lower: " 1-exp(-exp(r(estimate)-1.96*r(se))*10)
Lower: .04167308
. display "Upper: " 1-exp(-exp(r(estimate)+1.96*r(se))*10)
Upper: .05062594
```

Part 2

Question 3

(a)

We can read in the dataset, keep the rows required and `stset` the data for time to death, modelling for prostate cancer death. In our plot of the Kaplan-Meier curves, we restrict the y-axis using the `ylabel` option. From the first panel of the plot, we observe that the risk of prostate cancer death within ten years is very low for PSA values less than 3 ng/mL. Moreover, the risk for men with PSA values between 3 and 10 ng/mL is also comparatively low. The prostate cancer mortality risks for men with a PSA above 10 ng/mL are substantial, with approximately 10% of men dying due to prostate cancer by ten years. For men aged 60-69 years, there are only moderate competing risks, which are censored in these Kaplan-Meier curve calculations. Following a student's suggestion, we also plot the hazards (second panel). We find evidence that the hazards are rising with time, although we are cautious in our interpretation of the smoothed hazard curves.

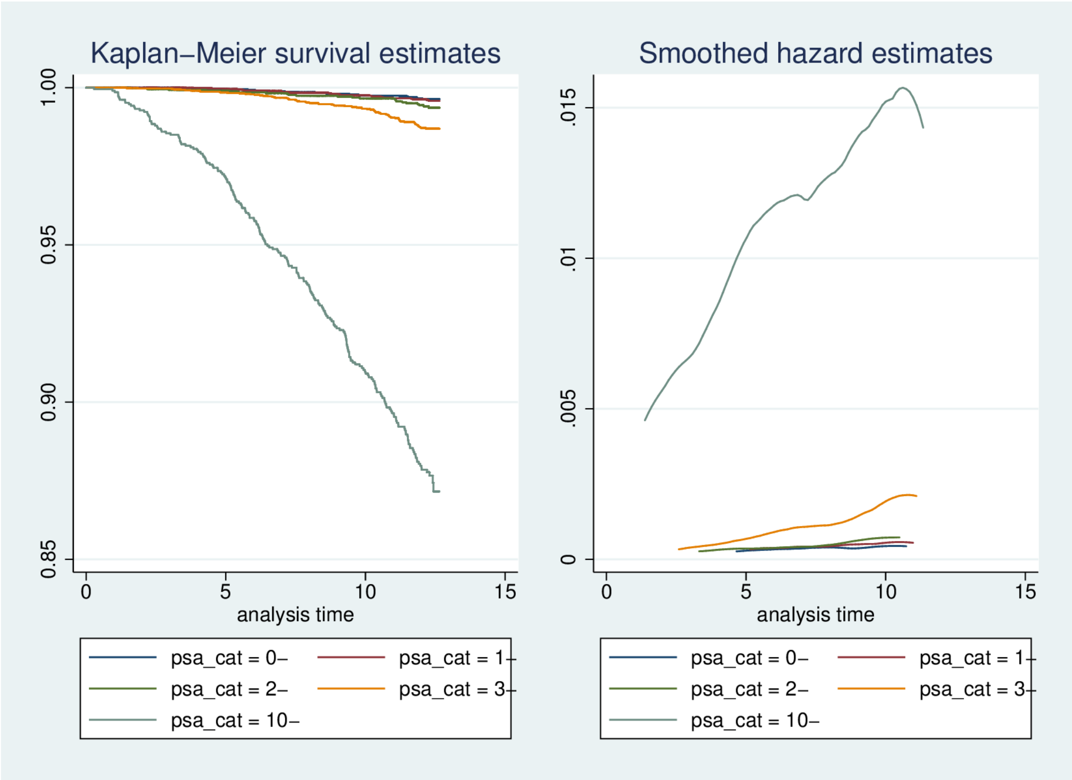
```
. use psa, clear
. keep if start_age>=60 & start_age<70
(68,607 observations deleted)
. stset age_dth, fail(event_dth==1) origin(start_age)

     failure event:  event_dth == 1
obs. time interval:  (origin, age_dth]
 exit on or before:  failure
    t for analysis:  (time-origin)
            origin:  time start_age

------------------------------------------------------------------------------
      31393  total observations
          0  exclusions
------------------------------------------------------------------------------
      31393  observations remaining, representing
        386  failures in single-record/single-failure data
 343413.911  total analysis time at risk and under observation
                                                at risk from t =         0
                                     earliest observed entry t =         0
                                          last observed exit t =  12.65369
. sts graph, by(psa_cat) ylabel(0.85(0.05)1) saving(q3a1, replace)

         failure _d:  event_dth == 1
   analysis time _t:  (age_dth-origin)
             origin:  time start_age
(file q3a1.gph saved)
. sts graph, by(psa_cat) hazard saving(q3a2, replace)

         failure _d:  event_dth == 1
   analysis time _t:  (age_dth-origin)
             origin:  time start_age
(file q3a2.gph saved)
. graph combine q3a1.gph q3a2.gph
. graph export q3a.eps, replace
(file q3a.eps written in EPS format)
. * the following line is only needed on Linux
. !! convert -density 300 q3a.eps q3a.png
```



We could further complement this analysis by a description of the level for the survival curves. For men with an initial PSA value that was 10 ng/mL or over, survival at five and ten years was 97.2% and 91.0%, respectively. For men with an initial PSA value between 3 and 10 ng/mL, ten-year survival was 99.3%.

```
. use psa, clear
. keep if start_age>=60 & start_age<70
(68,607 observations deleted)
. quietly stset age_dth, fail(event_dth==1) origin(start_age)
. sts list, by(psa_cat) at (5 10)

         failure _d:  event_dth == 1
   analysis time _t:  (age_dth-origin)
             origin:  time start_age

              Beg.                         Survivor      Std.
    Time     Total      Fail               Function     Error      [95% Conf. Int.]
-------------------------------------------------------------------------------
0-
       5      9245         5                0.9995    0.0002     0.9987    0.9998
      10      7890        17                0.9975    0.0005     0.9961    0.9983
1-
       5      7086         2                0.9997    0.0002     0.9989    0.9999
      10      6138        14                0.9976    0.0006     0.9961    0.9985
2-
       5      3913         4                0.9990    0.0005     0.9974    0.9996
      10      3417         9                0.9966    0.0010     0.9941    0.9980
3-
       5      6928        11                0.9985    0.0005     0.9972    0.9991
      10      6014        33                0.9933    0.0010     0.9911    0.9950
10-
       5      1908        58                0.9715    0.0037     0.9633    0.9779
      10      1552       114                0.9097    0.0066     0.8958    0.9217
-------------------------------------------------------------------------------
```

```
Note: Survivor function is calculated over full data and evaluated at
      indicated times; it is not calculated from aggregates shown at left.
```

An alternative approach would be to use life-tables.

```
. use psa, clear
. keep if start_age>=60 & start_age<70
(68,607 observations deleted)
. quietly stset age_dth, fail(event_dth==1) origin(start_age)
. ltable _t _d, by(psa_cat) interval(1(1)10)

                 Beg.                                       Std.
   Interval      Total   Deaths    Lost     Survival      Error     [95% Conf. Int.]
-------------------------------------------------------------------------------------
0-
    0     .      10072        0       5       1.0000     0.0000           .         .
    1     2      10067        0     176       1.0000     0.0000           .         .
    2     3       9891        0     173       1.0000     0.0000           .         .
    3     4       9718        2     212       0.9998     0.0001      0.9992    0.9999
    4     5       9504        3     257       0.9995     0.0002      0.9987    0.9998
    5     6       9244        1     267       0.9994     0.0003      0.9986    0.9997
    6     7       8976        5     249       0.9988     0.0004      0.9978    0.9993
    7     8       8722        2     252       0.9986     0.0004      0.9975    0.9992
    8     9       8468        4     283       0.9981     0.0005      0.9969    0.9988
    9    10       8181        5     287       0.9975     0.0005      0.9961    0.9983
   10     .       7889        8    7881       0.9954     0.0009      0.9933    0.9969
1-
    0     .       7666        0       3       1.0000     0.0000           .         .
    1     2       7663        0     117       1.0000     0.0000           .         .
    2     3       7546        0     149       1.0000     0.0000           .         .
    3     4       7397        1     146       0.9999     0.0001      0.9990    1.0000
    4     5       7250        1     164       0.9997     0.0002      0.9989    0.9999
    5     6       7085        4     179       0.9992     0.0003      0.9981    0.9996
    6     7       6902        1     174       0.9990     0.0004      0.9979    0.9995
    7     8       6727        4     188       0.9984     0.0005      0.9971    0.9991
    8     9       6535        3     206       0.9979     0.0006      0.9965    0.9988
    9    10       6326        2     187       0.9976     0.0006      0.9961    0.9985
   10     .       6137       10    6127       0.9944     0.0012      0.9915    0.9963
2-
    0     .       4173        0       2       1.0000     0.0000           .         .
    1     2       4171        1      49       0.9998     0.0002      0.9983    1.0000
    2     3       4121        2      52       0.9993     0.0004      0.9977    0.9998
    3     4       4067        0      70       0.9993     0.0004      0.9977    0.9998
    4     5       3997        1      84       0.9990     0.0005      0.9974    0.9996
    5     6       3912        2      98       0.9985     0.0006      0.9967    0.9993
    6     7       3812        1      81       0.9982     0.0007      0.9963    0.9992
    7     8       3730        3      96       0.9974     0.0008      0.9952    0.9986
    8     9       3631        0     101       0.9974     0.0008      0.9952    0.9986
    9    10       3530        3     111       0.9966     0.0010      0.9941    0.9980
   10     .       3416        9    3407       0.9913     0.0020      0.9864    0.9945
3-
    0     .       7386        0       4       1.0000     0.0000           .         .
    1     2       7382        2      89       0.9997     0.0002      0.9989    0.9999
    2     3       7291        0      95       0.9997     0.0002      0.9989    0.9999
    3     4       7196        5     117       0.9990     0.0004      0.9980    0.9995
    4     5       7074        4     143       0.9985     0.0005      0.9972    0.9991
    5     6       6927        4     181       0.9979     0.0005      0.9965    0.9987
    6     7       6742        7     160       0.9968     0.0007      0.9952    0.9979
    7     8       6575       11     174       0.9951     0.0008      0.9932    0.9965
```

```
    8     9       6390          5    198       0.9943     0.0009       0.9922     0.9959
    9    10       6187          6    168       0.9934     0.0010       0.9911     0.9951
   10     .       6013         37   5976       0.9812     0.0022       0.9763     0.9851
10-
    0     .       2096          2      2       0.9990     0.0007       0.9962     0.9998
    1     2       2092         13     22       0.9928     0.0019       0.9881     0.9957
    2     3       2057         15     33       0.9855     0.0026       0.9793     0.9898
    3     4       2009         11     34       0.9801     0.0031       0.9730     0.9853
    4     5       1964         17     40       0.9715     0.0037       0.9633     0.9779
    5     6       1907         25     40       0.9586     0.0044       0.9489     0.9665
    6     7       1842         23     42       0.9465     0.0051       0.9357     0.9556
    7     8       1777         18     45       0.9368     0.0055       0.9251     0.9467
    8     9       1714         24     53       0.9235     0.0061       0.9107     0.9345
    9    10       1637         24     62       0.9097     0.0066       0.8959     0.9218
   10     .       1551         55   1496       0.8474     0.0102       0.8262     0.8662
-------------------------------------------------------------------------------
```

(b)

We can compare the PSA categories using a log-rank test. For describing for the form of the association, we can use the interpretation from (a). We can also interpret the pattern of observed versus expected values, or, more directly, use Cox regression. From a Cox model with `pca_cat` as a categorical covariate, we find strong evidence for the change in mortality risk by PSA categories. Note that the Cox regression was not required here.

```
. sts test psa_cat

         failure _d:  event_dth == 1
   analysis time _t:  (age_dth-origin)
             origin:  time start_age


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

        |   Events         Events
psa_cat |  observed       expected
--------+-------------------------
0-      |        30         122.30
1-      |        26          94.54
2-      |        22          52.44
3-      |        81          92.19
10-     |       227          24.53
--------+-------------------------
Total   |       386         386.00

              chi2(4) =    1809.90
              Pr>chi2 =     0.0000
. stcox i.psa_cat, nolog

         failure _d:  event_dth == 1
   analysis time _t:  (age_dth-origin)
             origin:  time start_age

Cox regression -- no ties

No. of subjects =       31,393                  Number of obs    =      31,393
No. of failures =          386
Time at risk    =  343413.9114
```

```
                                                  LR chi2(4)      =     799.72
Log likelihood  =   -3528.6169                    Prob > chi2     =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     psa_cat |
         1-  |   1.120885   .3003368     0.43   0.670     .6629561    1.895122
         2-  |   1.709692   .4798985     1.91   0.056     .9862606    2.963768
         3-  |    3.58108   .7653739     5.97   0.000     2.355533     5.44426
        10-  |   37.74785    7.33327    18.69   0.000     25.79464    55.24017
------------------------------------------------------------------------------
```

(c)

We can use the `sts list` command to estimate the risks. The ten-year risks for those aged 60-69 years for PSA categories 0-, 1-, 2-, 3-9 and 10+ ng/mL were 0.25% (95% CI: 0.17, 0.39), 0.24% (95% CI: 0.15, 0.39), 0.34% (95% CI: 0.20, 0.59) and 9.0% (95% CI: 7.8, 10.4) respectively.

```
. sts list, by(psa_cat) at (5 10) fail

         failure _d:  event_dth == 1
   analysis time _t:  (age_dth-origin)
             origin:  time start_age

             Beg.                          Failure      Std.
    Time    Total     Fail                 Function     Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
0-
       5     9245        5                  0.0005     0.0002     0.0002    0.0013
      10     7890       17                  0.0025     0.0005     0.0017    0.0039
1-
       5     7086        2                  0.0003     0.0002     0.0001    0.0011
      10     6138       14                  0.0024     0.0006     0.0015    0.0039
2-
       5     3913        4                  0.0010     0.0005     0.0004    0.0026
      10     3417        9                  0.0034     0.0010     0.0020    0.0059
3-
       5     6928       11                  0.0015     0.0005     0.0009    0.0028
      10     6014       33                  0.0067     0.0010     0.0050    0.0089
10-
       5     1908       58                  0.0285     0.0037     0.0221    0.0367
      10     1552      114                  0.0903     0.0066     0.0783    0.1042
-------------------------------------------------------------------------------
Note: Failure function is calculated over full data and evaluated at indicated
      times; it is not calculated from aggregates shown at left.
```

A variety of other approaches would be used here, including life-tables, direct rate calculations and Poisson regression. Given evidence for a changing hazard, it would be sensible to split for time rather than assuming a constant hazard. Splitting at five years, we can use Poisson regression and `nlcom` using the following code:

```
. use psa, clear
. quietly keep if start_age>=60 & start_age<70
. quietly stset age_dth, fail(event_dth==1) origin(start_age) id(id)
. quietly stsplit fuband, at(0(5)15)
. streg i.fuband if psa_cat==4, dist(exp) nolog base

         failure _d:  event_dth == 1
```

```
   analysis time _t:  (age_dth-origin)
             origin:  time start_age
                 id:  id

Exponential regression -- log relative-hazard form

No. of subjects =        2,096                  Number of obs    =      5,554
No. of failures =          227
Time at risk    =  22183.74759
                                                LR chi2(2)       =      40.29
Log likelihood  =   -833.14411                  Prob > chi2      =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      fuband |
          0  |          1  (base)
          5  |   2.289127   .3692051     5.13   0.000     1.668717    3.140199
         10  |   2.820484   .5308447     5.51   0.000     1.950377    4.078765
             |
       _cons |   .0057385   .0007535   -39.30   0.000     .0044364    .0074227
------------------------------------------------------------------------------
. nlcom 1-exp(-(exp(_b[_cons])*5+exp(_b[_cons]+_b[5.fuband])*5))

       _nl_1:  1-exp(-(exp(_b[_cons])*5+exp(_b[_cons]+_b[5.fuband])*5))

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   .0900566   .0065639    13.72   0.000     .0771915    .1029216
------------------------------------------------------------------------------
```

Question 4

(a)

Fitting the Cox regression model with main effects for age and PSA, we find that age is strongly associated with prostate cancer mortality, with men aged 60-69 years having 7.38 (95% CI: 4.55, 11.97) times the prostate cancer mortality compared with those aged 50-59 years. For PSA categories, taking a reference group for those with an initial PSA of between 0 and 1 ng/mL, there was no significant difference in mortality for those with a PSA between 1 and 2 ng/mL (rate ratio (RR) = 1.12; 95% CI: 0.70, 1.80), although the number of events was small, as represented by the wider confidence intervals; for those with a PSA value between 2 and 3 ng/mL, the risk was higher (RR = 1.86, 95% CI: 1.14, 3.03).

```
. use psa, clear
. keep if start_age>=50 & start_age<70 & psa<3
(35,493 observations deleted)
. quietly stset age_dth, fail(event_dth==1) origin(start_age)
. drop if age_dth < start_age
(1 observation deleted)
. stcox i.age_cat i.psa_cat, nolog base

         failure _d:  event_dth == 1
   analysis time _t:  (age_dth-origin)
             origin:  time start_age

Cox regression -- no ties
```

```
No. of subjects =       64,506                  Number of obs    =     64,506
No. of failures =           99
Time at risk    =  725128.7179
                                                LR chi2(3)       =      95.81
Log likelihood  =   -1032.7654                  Prob > chi2      =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     age_cat |
        50-  |          1  (base)
        60-  |   7.379064   1.821191     8.10   0.000     4.549046    11.96967
             |
     psa_cat |
         0-  |          1  (base)
         1-  |   1.124633   .2685629     0.49   0.623     .7042771    1.795884
         2-  |   1.862285   .4630303     2.50   0.012     1.143951    3.031692
------------------------------------------------------------------------------
```

(b)

For the given dataset, there are several time scales of interest. First, we could consider time from the PSA test, which may be related to the time to the next PSA test, and hence to diagnosis, treatment and survival. However, for the men with PSA below 3 ng/mL, the time course to next PSA test is unclear. If we model for time since test, we then assume that the shape of the baseline hazard is the same for all groups. Calculations of survival and risks are simpler when we use this as the primary time scale. Second, we could adjust for attained age as the primary time scale. This is possibly a good choice, as age is closely related to prostate cancer mortality. The form of question (a) suggests that start age is categorical, but either time scale could be appropriate. We also provide code using attained age as the primary time scale. We see that the results comparing PSA categories are similar to those using time since test as the primary time scale. Third, in another dataset, we could possibly have used calendar period as the time scale, although changes in calendar period would generally be less than age or time since PSA test.

```
. stset age_dth, fail(event_dth==1) entry(start_age)

     failure event:  event_dth == 1
obs. time interval:  (0, age_dth]
 enter on or after:  time start_age
 exit on or before:  failure

------------------------------------------------------------------------------
      64506  total observations
          0  exclusions
------------------------------------------------------------------------------
      64506  observations remaining, representing
         99  failures in single-record/single-failure data
 725128.718  total analysis time at risk and under observation
                                                at risk from t =         0
                                     earliest observed entry t =        50
                                          last observed exit t =   81.6323
. stcox i.psa_cat, nolog base

         failure _d:  event_dth == 1
   analysis time _t:  age_dth
  enter on or after:  time start_age

Cox regression -- no ties
```

```
No. of subjects =       64,506                  Number of obs    =     64,506
No. of failures =           99
Time at risk    =  725128.7179
                                                LR chi2(2)       =       4.58
Log likelihood  =   -946.36737                  Prob > chi2      =     0.1014

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     psa_cat |
         0-  |          1  (base)
         1-  |   1.066148   .2545572     0.27   0.788     .6677001    1.702367
         2-  |   1.685212   .4187237     2.10   0.036     1.035516    2.742534
------------------------------------------------------------------------------
```

(c)

For the Cox model in 4 (a),

$$\lambda(t|age\_cat, psa\_cat) = \lambda_0(t)\exp(\beta_1 I(age\_cat = "60-") + \beta_2 I(psa\_cat = "1-") + \beta_3 I(psa\_cat = "2-"))$$

This uses the same notation as in 2 (c), with the extension that $\lambda_0(t)$ is the baseline hazard function for $psa\_cat = "0-"$ and $age\_cat = "50-"$.

(d)

$$Risk(t = 10|age\_cat = "60-", psa\_cat = "1-") = S_0(t = 10)^{HR(age=62,PSA=1.1)} = S_0(t = 10)^{\exp(\beta_1+\beta_2)}$$

where β_1 and β_2 are the regression parameters in 2 (c).

(e)

This answer assumes that 3 (c) was answered using the Kaplan-Meier estimator. The answer would change if another method had been used for 3 (c).

For 3 (c), survival and risks were calculated using the Kaplan-Meier estimator. For 4 (d), the risk is calculated from the Cox model, combining the hazard ratio with the Breslow estimator of baseline survival. These two approaches are closely related, where both assume a non-parametric baseline survival. The Kaplan-Meier curves are calculated separately for each stratum or group, with no modelling across strata or groups. The Cox model includes a model for the covariates, providing an opportunity to estimate risks for smaller groups under the assumption that the model holds. A Cox model stratified by both age and PSA categories would give the same as the Kaplan-Meier estimators.

(f)

As suggested by some of the students, we could first check whether there is any evidence for time-dependence using Schoenfeld residuals. These tests suggest no evidence for non-proportionality for either age or PSA categories.

```
. use psa, clear
. keep if start_age>=50 & start_age<70 & psa<3
(35,493 observations deleted)
. quietly stset age_dth, fail(event_dth==1) origin(start_age) id(id)
. quietly tab psa_cat, gen(psa_cat)
. quietly stcox i.age_cat i.psa_cat, nolog
. estat phtest, detail

      Test of proportional-hazards assumption

      Time:  Time
```

```
------------------------------------------------------------
            |       rho            chi2       df       Prob>chi2
------------+-----------------------------------------------
0b.age_cat  |          .              .        1             .
1.age_cat   |     -0.12656         1.60        1         0.2060
0b.psa_cat  |          .              .        1             .
1.psa_cat   |     -0.03177         0.10        1         0.7518
2.psa_cat   |     -0.03362         0.11        1         0.7364
------------+-----------------------------------------------
global test |                      1.84        3         0.6060
------------------------------------------------------------
```

Note that the numbers of prostate cancer deaths are not large, particularly in the first five years. Using (i) time-splitting at five years, we calculate indicators for splitting and use the indicators in the Cox model. The `stsplit` command takes care of the left truncation and event indicators. The hazard ratios for the time-varying effects are not individually significant and the Wald test also suggests no time-varying effects (p=0.83).

```
. use psa, clear
. keep if start_age>=50 & start_age<70 & psa<3
(35,493 observations deleted)
. stset age_dth, fail(event_dth==1) origin(start_age) id(id)

                id:  id
     failure event:  event_dth == 1
obs. time interval:  (age_dth[_n-1], age_dth]
 exit on or before:  failure
    t for analysis:  (time-origin)
            origin:  time start_age

------------------------------------------------------------------------------
      64507  total observations
          1  observation ends on or before enter()
------------------------------------------------------------------------------
      64506  observations remaining, representing
      64506  subjects
         99  failures in single-failure-per-subject data
 725128.718  total analysis time at risk and under observation
                                                at risk from t =         0
                                     earliest observed entry t =         0
                                          last observed exit t =  12.65369
. quietly tab psa_cat, gen(psa_cat)
. stsplit timeband, at(0,5,100)
(60,769 observations (episodes) created)
. gen timeband5 = timeband==5
. stcox i.age_cat i.psa_cat c.psa_cat2#c.timeband5 c.psa_cat3#c.timeband5, nolo
> g

         failure _d:  event_dth == 1
   analysis time _t:  (age_dth-origin)
             origin:  time start_age
                 id:  id

Cox regression -- no ties

No. of subjects =       64,506                  Number of obs    =     125,275
No. of failures =           99
Time at risk    =  725128.7179
                                                LR chi2(5)       =       96.19
```

```
Log likelihood  =   -1032.5763                 Prob > chi2      =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     age_cat |
        60-  |   7.379604   1.821329     8.10   0.000     4.549374    11.97056
             |
     psa_cat |
         1-  |   .8321468   .6077836    -0.25   0.801     .1988379    3.482578
         2-  |   2.109579   1.415804     1.11   0.266     .5661454    7.860746
             |
  c.psa_cat2#|
 c.timeband5 |   1.402602   1.083939     0.44   0.662     .3084046    6.378932
             |
  c.psa_cat3#|
 c.timeband5 |   .8659628   .6251042    -0.20   0.842     .2104024    3.564083
------------------------------------------------------------------------------
. testparm c.psa_cat2#c.timeband5 c.psa_cat3#c.timeband5

 ( 1)  c.psa_cat2#c.timeband5 = 0
 ( 2)  c.psa_cat3#c.timeband5 = 0

           chi2(  2) =    0.36
         Prob > chi2 =    0.8332
```

The code is simpler if we undertake a similar analysis using the `texp` and `tvc` options:

```
. use psa, clear
. keep if start_age>=50 & start_age<70 & psa<3
(35,493 observations deleted)
. quietly stset age_dth, fail(event_dth==1) origin(start_age) id(id)
. quietly tab psa_cat, gen(psa_cat)
. stcox i.age_cat i.psa_cat, nolog tvc(psa_cat2 psa_cat3) texp(_t >= 5)

         failure _d:  event_dth == 1
   analysis time _t:  (age_dth-origin)
             origin:  time start_age
                 id:  id


Cox regression -- no ties

No. of subjects =       64,506                 Number of obs    =     64,506
No. of failures =           99
Time at risk    =  725128.7179
                                               LR chi2(5)       =      96.19
Log likelihood  =   -1032.5763                 Prob > chi2      =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
main         |
     age_cat |
        60-  |   7.379604   1.821329     8.10   0.000     4.549374    11.97056
             |
     psa_cat |
         1-  |   .8321468   .6077836    -0.25   0.801     .1988379    3.482578
         2-  |   2.109579   1.415804     1.11   0.266     .5661454    7.860746
```

```
-------------+----------------------------------------------------------------
tvc          |
    psa_cat2 |   1.402602   1.083939     0.44   0.662     .3084046    6.378932
    psa_cat3 |   .8659628   .6251042    -0.20   0.842     .2104024    3.564083
------------------------------------------------------------------------------
Note: Variables in tvc equation interacted with _t>=5.
. test ([tvc]psa_cat2=0) ([tvc]psa_cat3=0)

 ( 1)  [tvc]psa_cat2 = 0
 ( 2)  [tvc]psa_cat3 = 0

           chi2(  2) =    0.36
         Prob > chi2 =    0.8332
```

For a continuous time-varying effect under (ii), we can also use the `texp` and `tvc` options. Again, there was no evidence for a time-varying effect.

```
. use psa, clear
. keep if start_age>=50 & start_age<70 & psa<3
(35,493 observations deleted)
. quietly stset age_dth, fail(event_dth==1) origin(start_age) id(id)
. quietly tab psa_cat, gen(psa_cat)
. stcox i.age_cat i.psa_cat, nolog tvc(psa_cat2 psa_cat3) texp(_t)

         failure _d:  event_dth == 1
   analysis time _t:  (age_dth-origin)
             origin:  time start_age
                 id:  id


Cox regression -- no ties

No. of subjects =       64,506                  Number of obs    =      64,506
No. of failures =           99
Time at risk    =  725128.7179
                                                LR chi2(5)       =       96.05
Log likelihood  =   -1032.6439                  Prob > chi2      =      0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
main         |
     age_cat |
        60-  |   7.381746    1.82195     8.10   0.000     4.550583    11.97433
             |
     psa_cat |
         1-  |   1.491879   1.166488     0.51   0.609     .3222445    6.906873
         2-  |   2.631034   2.124173     1.20   0.231     .5406374    12.80403
-------------+----------------------------------------------------------------
tvc          |
    psa_cat2 |   .9676291   .0839432    -0.38   0.704     .8163304    1.146969
    psa_cat3 |   .9604461   .0863422    -0.45   0.653     .8052889    1.145498
------------------------------------------------------------------------------
Note: Variables in tvc equation interacted with _t.
. test ([tvc]psa_cat2=0) ([tvc]psa_cat3=0)

 ( 1)  [tvc]psa_cat2 = 0
 ( 2)  [tvc]psa_cat3 = 0
```

```
           chi2(  2) =    0.24
         Prob > chi2 =    0.8862
```

Changing the time scale to attained age, we again found no evidence for a linearly time-varying effect (p=0.50).

```
. use psa, clear
. keep if start_age>=50 & start_age<70 & psa<3
(35,493 observations deleted)
. quietly stset age_dth, fail(event_dth==1) entry(start_age) id(id)
. quietly tab psa_cat, gen(psa_cat)
. stcox i.psa_cat, nolog tvc(psa_cat2 psa_cat3) texp(_t-60)

         failure _d:  event_dth == 1
   analysis time _t:  age_dth
  enter on or after:  time start_age
                 id:  id


Cox regression -- no ties

No. of subjects =       64,506                  Number of obs    =     64,506
No. of failures =           99
Time at risk    =  725128.7179
                                                LR chi2(4)       =       5.95
Log likelihood  =   -945.68365                  Prob > chi2      =     0.2033

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
main         |
     psa_cat |
         1-  |   1.447093   .9009586     0.59   0.553     .4271097    4.902907
         2-  |   3.375659   2.153459     1.91   0.057     .9668195    11.78614
-------------+----------------------------------------------------------------
tvc          |
    psa_cat2 |   .9743253   .0468119    -0.54   0.588     .8867632    1.070534
    psa_cat3 |   .9431796   .0471376    -1.17   0.242     .8551724    1.040244
------------------------------------------------------------------------------
Note: Variables in tvc equation interacted with _t-60.
. test ([tvc]psa_cat2=0) ([tvc]psa_cat3=0)

 ( 1)  [tvc]psa_cat2 = 0
 ( 2)  [tvc]psa_cat3 = 0

           chi2(  2) =    1.37
         Prob > chi2 =    0.5037
```

Finally, using stpm2, we again find no evidence of a time-varying effect.

```
. use psa, clear
. keep if start_age>=50 & start_age<70 & psa<3
(35,493 observations deleted)
. quietly stset age_dth, fail(event_dth==1) origin(start_age) id(id)
. quietly tab psa_cat, gen(psa_cat)
. stpm2 i.age_cat i.psa_cat, tvc(psa_cat2 psa_cat3) dftvc(2) df(3) scale(hazard
> ) nolog base

Log likelihood = -692.12018                     Number of obs     =     64,506
```

```
------------------------------------------------------------------------------
              |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+----------------------------------------------------------------
xb            |
      age_cat |
          50- |          0  (base)
          60- |   1.998929   .2468118     8.10   0.000     1.515186    2.482671
              |
      psa_cat |
           0- |          0  (base)
           1- |   .1301003    .261253     0.50   0.618    -.3819463    .6421468
           2- |   .7558834   .2635744     2.87   0.004     .2392871     1.27248
              |
        _rcs1 |   1.350218   .3070149     4.40   0.000     .7484803    1.951956
        _rcs2 |   .2412636   .1689149     1.43   0.153    -.0898034    .5723306
        _rcs3 |  -.0292311   .0230025    -1.27   0.204    -.0743153     .015853
_rcs_psa_c~21 |  -.0347741   .4192894    -0.08   0.934    -.8565663    .7870181
_rcs_psa_c~22 |   .0172805   .2135957     0.08   0.936    -.4013594    .4359204
_rcs_psa_c~31 |  -.5562371   .3425757    -1.62   0.104    -1.227673    .1151989
_rcs_psa_c~32 |   -.266076   .1840284    -1.45   0.148     -.626765     .094613
        _cons |  -8.081832   .2561783   -31.55   0.000    -8.583932   -7.579732
------------------------------------------------------------------------------
. testparm _rcs_psa*

 ( 1)  [xb]_rcs_psa_cat21 = 0
 ( 2)  [xb]_rcs_psa_cat22 = 0
 ( 3)  [xb]_rcs_psa_cat31 = 0
 ( 4)  [xb]_rcs_psa_cat32 = 0

           chi2(  4) =    4.22
         Prob > chi2 =    0.3773
```

(g)

Fitting a stratified Cox model with strata for PSA categories, we find evidence that the prostate cancer mortality rate is considerably higher in men aged 60-69 years compared with men aged 50-50 years (HR=7.38, 95% CI: 4.55, 11.97). The un-stratified Cox model includes one baseline hazard that is shared across the groups, while the stratified Cox model includes different baseline hazards for each stratum (in this case, PSA categories). This is a useful approach to deal with non-proportionality by PSA categories. However, given the lack of evidence for non-proportionality, the age effect is similar across the two models.

```
. use psa, clear
. keep if start_age>=50 & start_age<70 & psa<3
(35,493 observations deleted)
. quietly stset age_dth, fail(event_dth==1) origin(start_age) id(id)
. stcox i.age_cat, nolog strata(psa_cat) base

         failure _d:  event_dth == 1
   analysis time _t:  (age_dth-origin)
             origin:  time start_age
                 id:  id

Stratified Cox regr. -- no ties

No. of subjects =       64,506                  Number of obs    =     64,506
No. of failures =           99
Time at risk    =  725128.7179
```

```
                                                  LR chi2(1)      =      84.55
Log likelihood  =   -924.89875                    Prob > chi2     =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     age_cat |
        50-  |          1  (base)
        60-  |   7.381171   1.821808     8.10   0.000     4.550228     11.9734
------------------------------------------------------------------------------
                                                         Stratified by psa_cat
```

Question 5

(a)

We can define *safety* in terms of the levels of risk for prostate cancer incidence and prostate cancer death. As an additional analysis, we provide Kaplan-Meier estimators of the five- and ten-year risks for prostate cancer incidence and death by age and PSA categories (see Stata output and Table below). This is a simple, non-parametric approach that avoids modelling.

For men aged 50-59 years, the ten-year risk for prostate cancer incidence increases rapidly by PSA category, with approximately a 1% risk for PSA values below 1 ng/mL, 3% for PSA values between 1 and 2 ng/mL and 7% for PSA values between 2 and 3 ng/mL. The risks are considerably higher above 3 ng/mL because men are more likely to be referred to a urologist to undertake a biopsy to diagnose the cancer. For men in this age group with a PSA below 3 ng/mL, the ten-year risks of prostate cancer death are less than 0.1%. Interestingly, the ten-year risk of prostate cancer death among men with PSA between 3 and 10 ng/mL is only 0.3%, which is in marked contrast to the 19% ten-year risk of prostate cancer incidence. For men with a PSA value in excess of 10 ng/mL, the ten-year risk of prostate cancer death is 1 in 20.

For men aged 60-69 years at their initial PSA test, the risks are higher. The ten-year risks of prostate cancer diagnosis for men with PSA values between 0 and 1, 1 and 2, and 2 and 3 ng/mL are 2%, 4% and 9%, respectively; for the same PSA categories, the ten-year risks of prostate cancer death are 0.3, 0.2 and 0.3%, respectively.

The question arises as to the level of risk that is acceptable for choosing between 5 and 10 yearly re-testing. Is a ten-year risk of 0.3% acceptable for men in their 60s? As a working recommendation, the prostate cancer mortality risks are low, suggesting that 5 or 10 year re-testing would be 'safe'.

Note, however, that our interpretation of these data depend on the population, which has had moderately heavy testing. This issue is discussed further in the following question.

Age group (years)	PSA category (ng/mL)	Ten-year risk of PC incidence (%)	Ten-year risks of PC death (%)
50-59	0-	0.8	0.0
	1-	2.8	0.0
	2-	7.4	0.0
	3-	18.7	0.3
	10-	39.0	5.0
60-69	0-	2.3	0.3
	1-	4.3	0.2
	2-	9.4	0.3
	3-	22.8	0.7
	10-	42.0	9.0

```
. use psa, clear
. quietly stset age_dx, fail(event_dx==1) origin(start_age)
. sts list, by(age_cat psa_cat) at (5 10) fail

         failure _d:  event_dx == 1
   analysis time _t:  (age_dx-origin)
             origin:  time start_age
```

```
               Beg.                          Failure       Std.
    Time      Total     Fail                 Function      Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
50- 0-
       5      21966       67                  0.0029      0.0004      0.0023     0.0037
      10      19782      106                  0.0080      0.0006      0.0069     0.0093
50- 1-
       5      12918       61                  0.0046      0.0006      0.0036     0.0059
      10      11576      282                  0.0275      0.0015      0.0248     0.0305
50- 2-
       5       5394      131                  0.0233      0.0020      0.0196     0.0276
      10       4668      266                  0.0738      0.0036      0.0671     0.0811
50- 3-
       5       6260      536                  0.0773      0.0032      0.0713     0.0839
      10       5108      709                  0.1860      0.0048      0.1769     0.1956
50- 10-
       5       1219      245                  0.1638      0.0096      0.1459     0.1835
      10        818      317                  0.3897      0.0129      0.3650     0.4156
60- 0-
       5       9097      159                  0.0161      0.0013      0.0138     0.0188
      10       7711       62                  0.0234      0.0016      0.0206     0.0267
60- 1-
       5       7005       87                  0.0118      0.0013      0.0096     0.0146
      10       5886      204                  0.0427      0.0025      0.0381     0.0478
60- 2-
       5       3782      137                  0.0338      0.0028      0.0286     0.0398
      10       3096      222                  0.0944      0.0048      0.0855     0.1042
60- 3-
       5       6083      895                  0.1241      0.0039      0.1167     0.1320
      10       4655      677                  0.2280      0.0051      0.2182     0.2381
60- 10-
       5       1528      431                  0.2120      0.0091      0.1949     0.2305
      10        957      374                  0.4204      0.0114      0.3983     0.4431
70- 0-
       5       3289       94                  0.0250      0.0026      0.0205     0.0306
      10       2302       15                  0.0304      0.0029      0.0252     0.0366
70- 1-
       5       3028       31                  0.0094      0.0017      0.0067     0.0134
      10       2162       30                  0.0208      0.0027      0.0162     0.0267
70- 2-
       5       2025       43                  0.0194      0.0029      0.0144     0.0260
      10       1452       53                  0.0487      0.0049      0.0399     0.0592
70- 3-
       5       4483      373                  0.0720      0.0036      0.0653     0.0794
      10       3047      223                  0.1247      0.0048      0.1155     0.1345
70- 10-
       5       1737      386                  0.1663      0.0078      0.1517     0.1822
      10        933      247                  0.3068      0.0105      0.2868     0.3279
-------------------------------------------------------------------------------
Note: Failure function is calculated over full data and evaluated at indicated
      times; it is not calculated from aggregates shown at left.
. quietly stset age_dth, fail(event_dth==1) origin(start_age)
. sts list, by(age_cat psa_cat) at (5 10) fail

         failure _d:  event_dth == 1
   analysis time _t:  (age_dth-origin)
             origin:  time start_age
```

Time	Beg. Total	Fail	Failure Function	Std. Error	[95% Conf.	Int.]
50- 0-						
5	22030	0	0.0000	.	.	.
10	19939	4	0.0002	0.0001	0.0001	0.0005
50- 1-						
5	12978	1	0.0001	0.0001	0.0000	0.0005
10	11905	3	0.0003	0.0002	0.0001	0.0009
50- 2-						
5	5523	0	0.0000	.	.	.
10	5038	4	0.0008	0.0004	0.0003	0.0020
50- 3-						
5	6774	5	0.0007	0.0003	0.0003	0.0017
10	6251	16	0.0032	0.0007	0.0021	0.0049
50- 10-						
5	1443	20	0.0134	0.0030	0.0087	0.0207
10	1285	52	0.0502	0.0058	0.0400	0.0628
60- 0-						
5	9245	5	0.0005	0.0002	0.0002	0.0013
10	7890	17	0.0025	0.0005	0.0017	0.0039
60- 1-						
5	7086	2	0.0003	0.0002	0.0001	0.0011
10	6138	14	0.0024	0.0006	0.0015	0.0039
60- 2-						
5	3913	4	0.0010	0.0005	0.0004	0.0026
10	3417	9	0.0034	0.0010	0.0020	0.0059
60- 3-						
5	6928	11	0.0015	0.0005	0.0009	0.0028
10	6014	33	0.0067	0.0010	0.0050	0.0089
60- 10-						
5	1908	58	0.0285	0.0037	0.0221	0.0367
10	1552	114	0.0903	0.0066	0.0783	0.1042
70- 0-						
5	3376	16	0.0045	0.0011	0.0027	0.0073
10	2382	28	0.0140	0.0021	0.0104	0.0188
70- 1-						
5	3056	7	0.0022	0.0008	0.0010	0.0045
10	2206	15	0.0081	0.0017	0.0053	0.0122
70- 2-						
5	2061	9	0.0042	0.0014	0.0022	0.0080
10	1518	23	0.0170	0.0030	0.0120	0.0241
70- 3-						
5	4820	34	0.0067	0.0011	0.0048	0.0094
10	3463	106	0.0324	0.0027	0.0275	0.0381
70- 10-						
5	2071	172	0.0730	0.0054	0.0632	0.0843
10	1316	199	0.1768	0.0085	0.1609	0.1941

Note: Failure function is calculated over full data and evaluated at indicated times; it is not calculated from aggregates shown at left.

(b) The simplest answer here is that less testing would lead to fewer prostate cancer diagnoses, where men with prostate cancer but without clinical symptoms would die due to other causes. Based on the European Randomised Study of Prostate Cancer, we could expect that mortality would increase by approximately 20% with no testing and prostate cancer incidence would be considerably lower.

The risks from (a) are estimated from a population with moderately intense PSA testing. The research question relates to the counterfactual *were a man to not be tested, what would be the ten-year risks?* Under that counterfactual, we have under-estimated the risks of death by approximately 20%.

Various answers discussed the issue of *lead-time bias*, which is more of an issue with survival from prostate cancer incidence to death, rather than mortality rates from an cohort with no previous diagnosis of prostate cancer.

Question 6

The partial likelihood for a Cox regression model with a single covariate is

$$L = \prod_i \frac{\exp(\beta x_i)}{\sum_{j \in R_i} \exp(\beta x_i)}$$

where i is an index for the events, j is an index for the risk set R_i for event i, and x_i and x_j are covariates.

For a nested case-control study, the likelihood for a single covariate is

$$L = \prod_i \frac{\exp(\beta x_i)}{\sum_{j \in R_i^*} \exp(\beta x_i)}$$

where i is an index for the events (or cases), j is an index for a *sample* of the risk R_i^* (or controls) for event (or case) i, and x_i and x_j are the covariates.

The difference in the formulations is the risk set: the full risk set is used for Cox regression, while the nested case-control study only includes a sample of the risk set. The sampling from the risk set will decrease the precision of the estimated regression parameters.

The odds ratio from the nested case-control study will estimate a hazard ratio.