Biostat3: Risk estimation

Mark Clements

March 15, 2017

In outline, we consider various approaches to estimate 5-year risks of colon cancer death for females following a colon cancer diagnosis by stage of disease.

We read in the dataset:

```
. set linesize 80
. cd /home/marcle/repos/biostat3_2014/tmp/
/home/marcle/repos/biostat3_2014/tmp
. use colon, clear
(Colon carcinoma, all stages, 1975-94, follow-up to 1995)
```

We then stset the data (after creating the `id` variable):

```
. capture drop id
. gen id = _n
. stset exit, fail(status==1) origin(dx) id(id) scale(365.24)

                id:  id
     failure event:  status == 1
obs. time interval:  (exit[_n-1], exit]
 exit on or before:  failure
    t for analysis:  (time-origin)/365.24
            origin:  time dx

--------------------------------------------------------------------------------
      15564  total observations
          0  exclusions
--------------------------------------------------------------------------------
      15564  observations remaining, representing
      15564  subjects
       8369  failures in single-failure-per-subject data
  58546.178  total analysis time at risk and under observation
                                                at risk from t =         0
                                     earliest observed entry t =         0
                                          last observed exit t =  20.98073
```

As a first approach, we can calculate non-parametric risk (or failure) estimates from the Kaplan-Meier curves using the `sts list` command:

```
. sts list if agegrp==2 & stage!=0 & sex==2, by(stage) at(5 10) failure

         failure _d:  status == 1
   analysis time _t:  (exit-origin)/365.24
             origin:  time dx
                 id:  id

              Beg.                        Failure       Std.
    Time      Total      Fail             Function      Error     [95% Conf. Int.]
--------------------------------------------------------------------------------
```

```
Localised
       5       793      310              0.2317    0.0117     0.2097    0.2556
      10       338       52              0.2956    0.0138     0.2694    0.3237
Regional
       5       136      225              0.5299    0.0250     0.4819    0.5796
      10        50       18              0.6109    0.0274     0.5577    0.6646
Distant
       5        87     1102              0.9072    0.0088     0.8890    0.9235
      10        39       15              0.9276    0.0084     0.9101    0.9429
------------------------------------------------------------------------------
Note: Failure function is calculated over full data and evaluated at indicated
      times; it is not calculated from aggregates shown at left.
```

As a second approach, we can fit a Poisson regression with constant rates:

```
. streg i.sex##i.stage if agegrp==2 & stage!=0, dist(exp) nolog base

         failure _d:  status == 1
   analysis time _t:  (exit-origin)/365.24
             origin:  time dx
                 id:  id

Exponential regression -- log relative-hazard form

No. of subjects =        5,735                 Number of obs    =       5,735
No. of failures =        3,087
Time at risk    =  22828.45526
                                               LR chi2(5)       =     4300.57
Log likelihood  =   -8039.8071                 Prob > chi2      =      0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
        Male |          1  (base)
      Female |   .8576709    .065315    -2.02   0.044     .7387515    .9957332
             |
       stage |
  Localised  |          1  (base)
   Regional  |   2.642185   .2632503     9.75   0.000     2.173477     3.21197
    Distant  |   13.62924   .9005805    39.53   0.000     11.97365    15.51374
             |
   sex#stage |
     Female #|
   Regional  |   1.258471   .1618644     1.79   0.074     .9780514     1.61929
     Female #|
    Distant  |   1.070361   .0947776     0.77   0.443     .8998258    1.273215
             |
       _cons |   .0463934   .0026307   -54.15   0.000     .0415135     .051847
------------------------------------------------------------------------------
```

We can obtain confidence intervals for the rates using the `lincom` command:

```
. lincom _cons+2.sex, eform

 ( 1)  [_t]2.sex + [_t]_cons = 0

------------------------------------------------------------------------------
          _t |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
```

```
-------------+------------------------------------------------------------------
         (1) |   .0397903   .0020227   -63.43   0.000     .0360171    .0439588
------------------------------------------------------------------------------
```

Using the confidence intervals for the rate, we can calculate the five-year risk and its 95% confidence intervals:

```
. di 1-exp(-.0397903*5)
.18041036
. di 1-exp(-.0360171*5)
.1648012
. di 1-exp(-.0439588*5)
.19731587
```

Finally, we can use Cox regression:

```
. stcox i.sex##i.stage if agegrp==2 & stage!=0, nolog base

        failure _d:  status == 1
  analysis time _t:  (exit-origin)/365.24
            origin:  time dx
                id:  id

Cox regression -- Breslow method for ties

No. of subjects =        5,735                 Number of obs    =       5,735
No. of failures =        3,087
Time at risk    =  22828.45526
                                               LR chi2(5)       =     2775.22
Log likelihood  =    -23864.82                 Prob > chi2      =      0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
        Male |          1  (base)
      Female |   .9418098   .0717916    -0.79   0.432     .8111074    1.093574
             |
       stage |
  Localised  |          1  (base)
   Regional  |   2.285848   .2278625     8.29   0.000     1.880166    2.779064
    Distant  |   8.842267   .5936013    32.47   0.000     7.752121    10.08572
             |
   sex#stage |
     Female #|
   Regional  |   1.179394   .1517318     1.28   0.200     .9165369    1.517637
     Female #|
    Distant  |   1.053203   .0932987     0.59   0.558      .885335      1.2529
------------------------------------------------------------------------------
```

We can pull out $S_0(5)$ using the `predict` function with the `basesurv` argument. One way to get the times is the following:

```
. preserve
. sort agegrp stage _t
. predict S0, basesurv
(9,829 missing values generated)
. keep if agegrp==2 & stage!=0 & _t<5
(11,506 observations deleted)
```

```
. by agegrp stage: list _t S0 if _n==_N

-------------------------------------------------------------------------------
-> agegrp = 60-74, stage = Localised

      +----------------------+
      |        _t          S0 |
      |----------------------|
1408. | 4.9802869   .74343243 |
      +----------------------+

-------------------------------------------------------------------------------
-> agegrp = 60-74, stage = Regional

      +----------------------+
      |        _t          S0 |
      |----------------------|
 545. | 4.8954112   .74512036 |
      +----------------------+

-------------------------------------------------------------------------------
-> agegrp = 60-74, stage = Distant

      +----------------------+
      |        _t          S0 |
      |----------------------|
2105. | 4.9583835   .74343243 |
      +----------------------+
. restore
```

We can combine the baseline survival with the hazard ratio to calculate the risk:

```
. display 1-.74343243^.9418098
.24363054
```

Confidence interval estimation can be done with the use of the `survi` user-contributed command:

```
. survci if agegrp==2 & stage!=0, survival at(sex=2, stage=1) outfile(pred, rep
> lace)
(sex=2.00; stage=1.00)

file pred.dta saved
. preserve
. use pred, clear
(estimates adjusted for sex=2; stage=1)
. sort _t
. keep if _t<5
(1,677 observations deleted)
. gen _fail = 1-_surv
. gen _fail_lb = 1-_ub
. gen _fail_ub = 1-_lb
. list _t _fail _fail_lb _fail_ub if _n==_N

      +-------------------------------------------------+
      |        _t       _fail    _fail_lb    _fail_ub |
      |-------------------------------------------------|
4058. | 4.9802869   .24363053   .22301758   .26580314 |
      +-------------------------------------------------+
. restore
```