A discrete-time split population survival ('cure') model for Stata 6 or Stata 7: `spsurv`

Stephen P. Jenkins

Institute for Social and Economic Research

University of Essex

Email: stephenj@essex.ac.uk

Outline

- The model
- **`spsurv`** syntax
- Illustration using the cancer data (**`cancer.dta`**)
- Reflections
  - heterogeneity in the cure probability?
  - maximisation issues (‘backing up’)
  - `robust` option is infeasible

Thanks to StataCorp Technical Support for their help!

Overview

- **`spsurv`** estimates what economists refer to as split population survival models (Schmidt and Witte, 1989) and biostatisticians refer to as cure models, for the case where
- survival time metric is intrinsically discrete or survival times are grouped into intervals.
- Cf. the continuous time lognormal cure model **`lncure`** by Mario Cleves (**`st`** compatible, most **`streg`** features and options, but particular parametric hazard shape)
- **`ml`**, method **`d0`** (can't use **`lf`**)

The model

- Standard survival models assume that prob(eventual failure) > 0 for all individuals; split population models suppose that a proportion, *c*, never fail ('cured').
- Likelihood contribution for person *i* with survival time *t*:

$$\ln L_i = d_i.\ln[(1\text{-}c).(h_{it}).(S_{it\text{-}1})] + (1\text{-}d_i)\ln[c + (1\text{-}c).S_{it}]$$

where d_i is a binary censoring indicator (=1 if failure, 0 if right-censored), S_{it} is the discrete-time survivor function, and the (cloglog) discrete-time hazard rate

$$h_{it} = 1 - \exp[-\exp(I_{it})]; \qquad I_{it} = f(t) + b'X_{it}$$

spsurv depvar varlist [**if** <exp>] [**in** <range>] , **id(**idvar**) seq(**seqvar**)** [**nocons**] [**cpr0(#) eform level(#)** **mlopts**]

- Data organised in person-month form (**expand**)
- depvar event indicator in each period at risk of event (derive from censoring indicator)
- varlist covariates, including duration dependence
- idvar person identifier
- seqvar spell interval identifier for each *i* (1,…,*t*)
- **cpr0(#)** value of logit(c) used as starting value (default = -4, i.e. a cure probability of about 0.018)

Illustration (i): set up the data

```
. use cancer
(Patient Survival in Drug Trial)
. ge id = _n  /* create unique person identifier */
. expand studytim /* 1 obs/month at risk of death */
(696 observations created)
. sort id
. quietly by id: ge t = _n /*spell month id, by i */
. quietly by id: ge dead = died & _n==_N  /* depvar */
. * drug = 1 (placebo); drug =2,3 (receives drug)
. recode drug 1=0 2/3=1
(744 changes made)
. lab var drug "1=receives,0=placebo"
. ge logt = ln(t) /* duration dependence */
```

Illustration (ii): **`cloglog`** model, used to derive starting values

```
. cloglog dead drug age logt

Iteration 0:   log likelihood =  -111.3772
Iteration 1:   log likelihood =   -111.264
Iteration 2:   log likelihood = -111.26371
Iteration 3:   log likelihood = -111.26371

Complementary log-log regression                Number of obs     =        744
                                                Zero outcomes     =        713
                                                Nonzero outcomes  =         31

                                                LR chi2(3)        =      35.20
Log likelihood = -111.26371                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
        dead |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
        drug |   -2.18907   .4110876     -5.325   0.000      -2.994787   -1.383353
         age |    .119348   .0371648      3.211   0.001       .0465064    .1921896
        logt |   .6402733   .2454492      2.609   0.009       .1592017    1.121345
       _cons |  -9.928747   2.272995     -4.368   0.000      -14.38374   -5.473759
------------------------------------------------------------------------------
```

Illustration (iii): **spsurv**

```
. spsurv  dead drug age logt, id(id) seq(t)

Iteration 0:   log likelihood = -111.60074
Iteration 1:   log likelihood = -111.26779
<snip>
Iteration 5:   log likelihood = -111.26372
Iteration 6:   log likelihood = -111.26371

Split population survival model                  Number of obs   =        744
                                                 LR chi2(4)      =      35.20
Log likelihood = -111.26371                      Prob > chi2     =     0.0000

------------------------------------------------------------------------------
    dead |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
hazard   |
    drug |  -2.189079   .4110974     -5.325   0.000      -2.994815   -1.383343
     age |   .1193277    .037166      3.211   0.001       .0464837    .1921717
    logt |   .6401813   .2454488      2.608   0.009       .1591105    1.121252
   _cons |  -9.927432   2.273042     -4.367   0.000      -14.38251    -5.47235
---------+--------------------------------------------------------------------
cure_p   |
   _cons |  -16.43746   325.1069     -0.051   0.960      -653.6352    620.7603
------------------------------------------------------------------------------
Pr(never fail) = 7.266e-08; Std.Err. = .00002362; z = .00307591; P>|z| = 0.998
```

Other issues

- Heterogeneity in the cure probability, *c*?
  - OK to program, but hard to derive signif. estimates
- 'Backing up' in maximization with some test data sets ('true' maximum overshot)
- S_E globals get zapped by **`ml`** in version 6 but not version 5 or 7!
- Robust option -- requires d1 -- 'infeasible' (true for other programs with data in groups): likelihood not of the linear form such that can take derivative w.r.t. to Xbeta (1 score vector per equation). Harder than Gould/Sribney ML book examples might suggest!