Expected and Relative Survival

Vincenzo Coviello
Department of Prevention ASL BA/1
Minervino Murge (Ba)
Email: coviello@mythnet.it

Outline of talk

- Estimating Expected Survival
- stexpect
- Example 1: clinical survival study
- Example 2: Population-based survival study

ESTIMATING EXPECTED

SURVIVAL (1)

Definition

Expected survival is the survival in a reference population which is similar to the study cohort of patients at the start of follow-up, where the matching factors are usually age, calendar time, sex and optionally other variables (race, census).

The estimate is achieved through population mortality rate tables.

Using population mortality rates:

stexpect

- Estimates individual expected survival, the building block of the overall curve.

- Combines these individual estimates to give the expected survival of the cohort according to three methods:
    - Ederer or "*exact*"
    - Hakulinen
    - Conditional or Ederer II

Individual Expected Survival

- A 36 years old man born on 23th April 1964
- Followed-up from 15th June 2000 to 25th October 2001

Follow-up		Hazard per day	Cumulative hazard (Λ)	Survival probability exp(-λ)
From	To			
15-Jun-2000	23-Apr-2001	0.00000155	0.0004836	0.999516517
23-Apr-2001	25-Oct-2001	0.00000161	0.00029785	0.999702194
	Cumulative hazard from 15 Jun-2000 to 25-Oct-2001 =		0.00078145	0.999218855

Formulas

- Ederer and Hakulinen method:

$$S_e(t+s) = S_e(t)\frac{\sum S_i(t+s)}{\sum S_i}$$

- Conditional or Ederer 2 method:

$$S_e(t+s) = S_e(t)\exp\frac{\sum h_i(t,s)Y_i(t)}{\sum Y_i(t)}$$

where Y_i is 1 if the subject is at risk at time *t* and 0 otherwise.

Problems in large data sets

- To compute the above equations the time axis must be partitioned at every observed failure and censored time.
- In large data sets this episode splitting may require huge amounts of memory.

Approximation

- The range of follow-up times is partitioned in *n* evenly spaced points. In such fixed width intervals each subject will contribute to the expected survival with a weight equals to the proportion of time for which he is observed.

Ederer - Hakulinen approximate formula :

$$S_e(t+s) = S_e(t)\,\frac{\sum S_i w_i(t+s)}{\sum S_i w_i}$$

where $w_i = (t_i - t)/s$

stexpect

stexpect …, ratevar(varname) output(filename [,replace]) [method(#)]

They are not options

- **ratevar(varname)** : variable containing the general population mortality rates

output(filename [,replace]) : file where the estimates will be stored

method(#) : methods to be used

1 = Ederer I

2 = Ederer II or Conditional

3 = Hakulinen (default)

stexpect …, … [by(varlist) at(numlist) np(#)]

by(varlist) : up to 5 variables specifying separate groups over which the expected survival is to be calculated.

at(numlist) : analysis times at which the expected survival is to be computed.

npoints(#) : number of equally spaced points in the range of follow-up times used for the approximate estimate.

Before using stexpect one needs to

- stset data using the id() option.
- split follow-up time by age and calendar period.
- merge the cohort data set with the file of reference population rates.

Example 1

Clinical Survival Study

MGUS Study

- 241 cases of Monoclonal Gammopathy of Undetermined Significance.
- **time** is in days since identification to death or occurrence of lymphoproliferative disease or to the end of the study.
- **status** is a failure/censor indicator.

```
Contains data from C:\Convegni2004\mgusconvegno.dta
  obs:           241
 vars:            12                          11 Aug 2004 07:13
------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------
id              int    %9.0g
sex             byte   %9.0g
time            float  %9.0g                  Time since Diagnosis
status          byte   %17.0g      status
…omitted…
```

Preparing the dataset

1 – stset data

```
. stset time, f(status) id(id) scale(365.25)
```

2 – split the follow-up time by age and calendar period

```
. stsplit fu, at(0(1)25)

. gen age = agedia+fu

. gen year = yeardiagnosis + fu
```

3 – merge the cohort data with a file (usrate) of reference rates

```
. sort year age sex

. merge year age sex using usrate, keep(rate) uniqus nokeep
```

```
texpect, ratevar(rate) out(cond_example,replace
                     method(2)
```

- rate is the variable containing reference population rates
- method(2) specifies that the conditional estimate is to be computed
- cond_example is the output file structured as follows:

```
. use cond_example,clear
. list in 1/3, noobs

              +----------------------------------+
              |      t_exp    atrisk      Survexp |
              |----------------------------------|
              |          0       241            1 |
              |  .00027405       241    .99998966 |
              |  .08487337       239     .9968664 |
              +----------------------------------+
```

Output file

Survexp saves the estimate of the expected survival. The user can define a different name for this variable:

`stexpect [  newvarname ] ,...`

t_exp stores the times at which the function is estimated. If at(numlist) is omitted, they correspond to each survival time.

atrisk contains the number of subjects at risk at the time t_exp.

Check the validity of the results

The table below lists the results at the last five follow-up times achieved by stexpect and by the R macro survexp.

```
. list t_exp  Survexp R_est in -5/l,noobs

        +--------------------------------+
        |    t_exp      Survexp    R_est |
        |--------------------------------|
        | 26.277892   .22859971    .2286 |
        | 27.359343      .20821    .2082 |
        | 27.712526   .20168448    .2017 |
        | 28.361396   .18769732    .1877 |
        | 34.105407   .07531006    .0753 |
        +--------------------------------+
```

at(numlist) and by(varlist)

To illustrate these options new conditional estimates are saved in the file cond_byex :

```
stexpect,ratevar(rate) out(cond_byex,replace) ///
            method(2) at(0(1)25) by(sex)
```

The file cond_byex will record the expected survival

- at the times t_exp = 0 , 1 , 2 , …. , 24 , 25
- for each value of byvar sex .

Output file with **by(varlist)** and **at(numlist)**

```
. use cond_byex,clear

. list if t_exp>20,noobs


      +----------------------------------+
      | sex   t_exp    atrisk    Survexp |
      |----------------------------------|
      |   1      21        19   .24535683 |
      |   1      22        11   .22539159 |
      |   1      23         8    .2075762 |
      |   1      24         6   .18930929 |
      |   1      25         4   .17506198 |
      |----------------------------------|
      |   2      21        21   .45990346 |
      |   2      22        12     .434795 |
      |   2      23         7   .40512875 |
      |   2      24         5   .38333169 |
      |   2      25         4   .36152862 |
      +----------------------------------+
```

Other methods

- To estimate the expected survival according to Ederer or Hakulinen method, the follow-up time of the subjects must be set differently.

- So the expected survival of the three methods cannot be estimated sequentially, because each of them needs a different timevar in the stset statement.

Some Comment

- To estimate the expected survival, subjects in data set are to be considered as elements within the reference population. Fixing the follow-up of these elements at the observed times in the study cohort, as in Conditional method, is meaningless.

- Follow-up time in Ederer and Hakulinen methods actually matches the expected survival definition “The survival in a reference population which is similar to the study cohort of patients **at the start of follow-up**”.

Follow-up Time

- **Ederer Method**

  The follow-up time is the same for all of the subjects and corresponds to the largest time at which an expected survival estimate is required.

- **Hakulinen Method**

  The follow-up time is the actual censoring time for those subjects who are censored and the “maximum potential follow-up” for those who have died.

  Find the rationale in references (3) and (4).

Ederer Method

Expected Survival until 25 years from diagnosis

1 – stset

```
gen survederer =  25*365.25

stset  survederer, f(status) id(id) scale(365.25)
```

2 – merge with the file of reference rates

```
stsplit fu,at(0(1)35)
gen age = aged+fu
gen year=yeard+fu
sort year age sex
merge year age sex using c:\data\usrate, nokeep ///
keep(rate)
```

Ederer Method with stexpect

```
stexpect,ratevar(rate) out(ederer_ex,replace) ///
      method(1) at(0(1)25) by(sex)
```

- method(1) tells stexpext to use the Ederer-Hakulinen formula.

- at(numlist) is not an option in this method since no failure occurs during the follow-up.

Results with Ederer Method

```
. use ederer_ex,clear

. list if t_exp<5, noobs

 +--------------------------------+
 | sex   t_exp   atrisk   Survexp |
 |--------------------------------|
 |   1       0      140         1 |
 |   1       1      140  .95254107 |
 |   1       2      140  .90595187 |
 |   1       3      140  .86049999 |
 |   1       4      140  .81635571 |
 |--------------------------------|
 |   2       0      101         1 |
 |   2       1      101  .97917553 |
 |   2       2      101  .95746886 |
 |   2       3      101  .93495095 |
 |   2       4      101  .91170243 |
 +--------------------------------+
```

Note that the number at risk does not change.

Hakulinen’s Method

The “maximum potential follow-up time” for failed subjects is settled as the difference between the most optimistic last contact date and the enrollment date.

The MGUS study ends at August 1, 1990. So, the survival time according to Hakulinen’s method is set as:

```
gen survhakulinen = cond(status,mdy(8,1,1990)-datediag,time)

stset  survhakulinen, f(status) id(id) scale(365.25)
```

Merge instructions are omitted.

Hakulinen’s Method with stexpect

```
stexpect,ratevar(rate) out(hakulinen_ex,replace) ///
                           at(0(1)25) by(sex)
```

method(3) is omitted because it is the default.

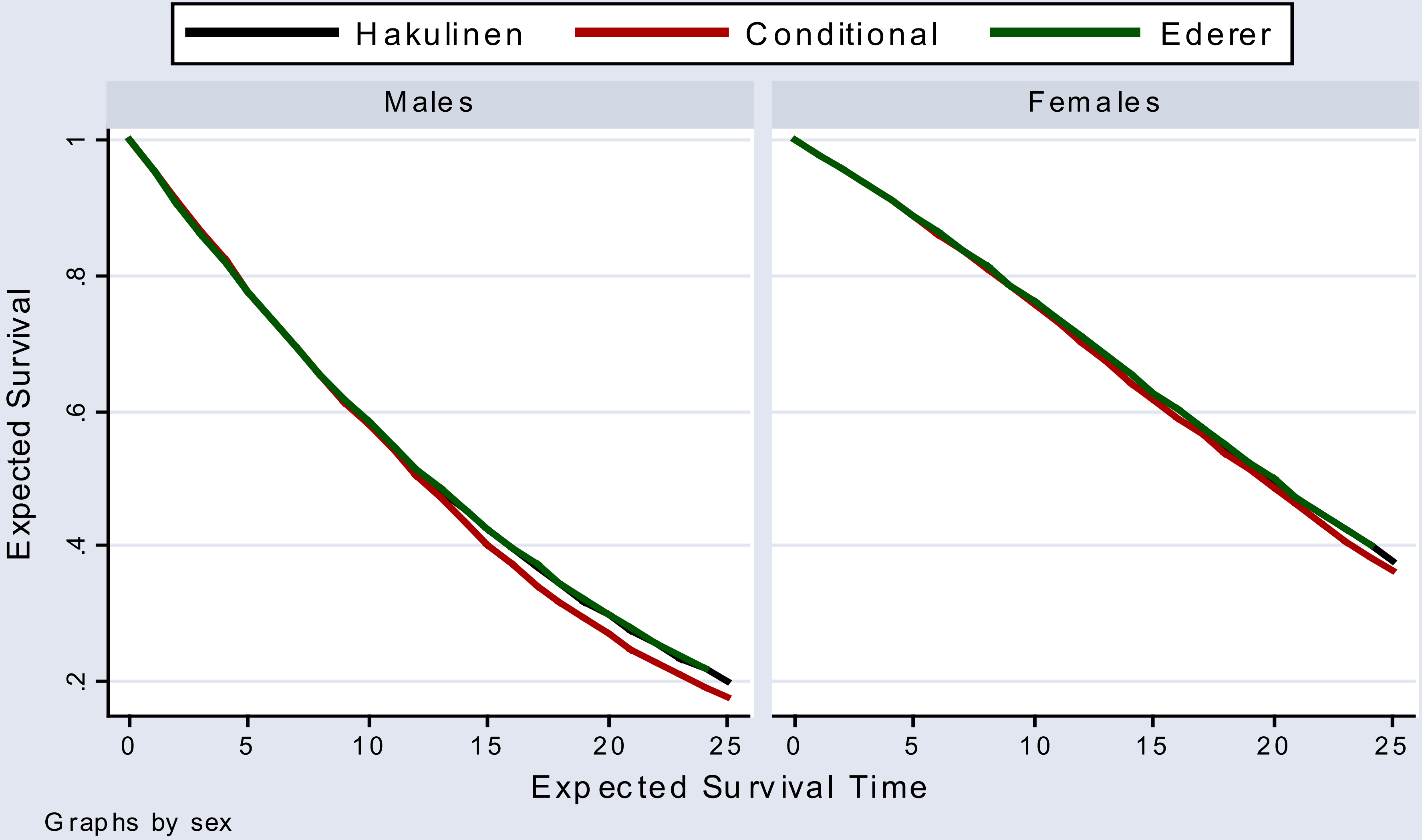
Since the follow-up time has been modified, the number of subjects at risk is not the same as in the study cohort.

```
+----------------------------------+
|    sex   t_exp   atrisk   Survexp |
|----------------------------------|
|   Males      7      140   .6924624 |
|   Males      8      138  .65401214 |
|   Males      9      138  .61671766 |
|   Males     10      138  .58095552 |
|   Males     11      138  .54669681 |
|   Males     12      138  .51385414 |
|----------------------------------|
| Females      7      101  .83828516 |
| Females      8      101  .81280358 |
| Females      9      101  .78695202 |
| Females     10      101  .76083501 |
| Females     11      100  .73444907 |
| Females     12      100  .70776837 |
+----------------------------------+
```

Comparison of Methods

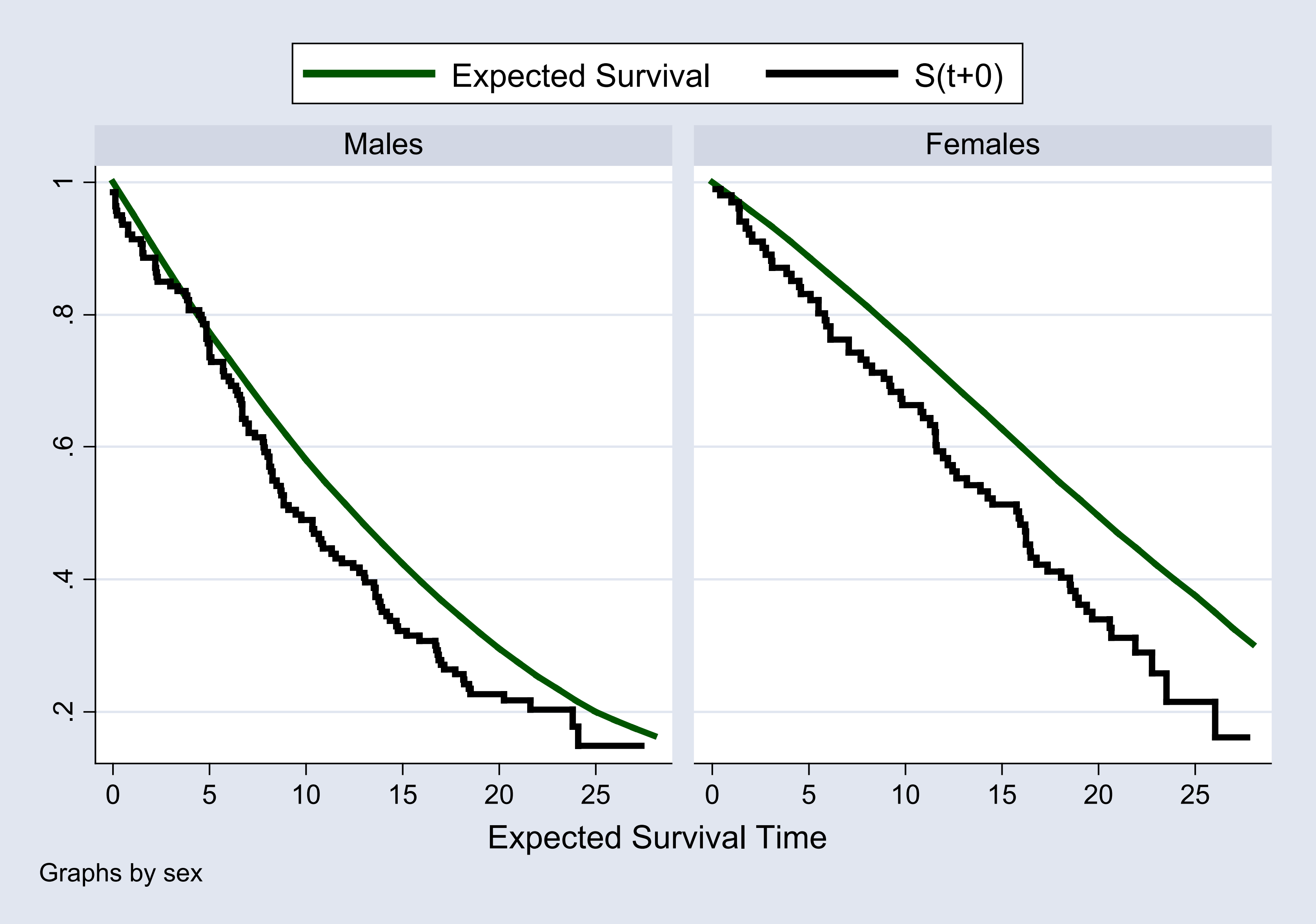
In the next slide a graph comparing the three estimates is shown. Here are the lines to achieve it:

```
use hakulinen_ex,clear
rename Survexp Hakulinen
merge sex t_exp using ederer_ex, keep(Survexp)
rename Survexp Ederer
drop _m
sort sex t_exp
merge sex t_exp using cond_byexample, keep(Survexp)
rename Survexp Conditional
twoway line Hakulinen Conditional Ederer t_exp, ///
   legend(label(1 "Hakulinen") label(2 "Conditional") ///
   label(3 "Ederer") row(1)) xla(0(5)25) ///
   by(sex, legend(pos(12))) clc(black red green)
```



- he three methods often yield similar results.
- he Conditional estimate is of a small amount lower than th
derer and Hakulinen estimates, which overlap completely

Comparison of Observed vs. Expected Survival



Example 2

Population-based Survival Study

Relative Survival

- Relative survival is the preferred measure for survival analysis based on population cancer registry data mainly because it does not depend on the information on cause of death.

- It is computed as the ratio between observed and expected survival.

- Relative survival can be estimated using sts gen to produce an estimate of the observed survival and stexpect for the expected survival.

Melanoma Data of the Finnish Cancer Registry

2145 patients with localised skin melanoma in Finland during 1975-1984.

```
. use melanoma,clear
(Skin melanoma, all stages, Finland 1975-94, follow-up to 1995)

. keep if year8594==0 & stage==1

Contains data from melanoma.dta
  obs:          2,145
 vars:             14                            11 Aug 2004 18:14
-------------------------------------------------------------------
              storage  display     value
variable name   type   format      label   variable label
-------------------------------------------------------------------
id              int    %9.0g
sex             byte   %9.0g       sex     Sex
surv_mm         float  %9.0g               Survival time in
                                           completed months
status          byte   %17.0g      status  Vital status at last
                                           date of contact
…omissis
```

Hakulinen's Method for Relative Survival

- surv_mm is the timevar in months from diagnosis,
- status is coded 1 or 2 if death occurs and 0 otherwise.

The analysis cutoff is set at December 31, 1995

As shown before, the survival time must be adapted to get the Hakulinen expected survival:

```
.gen surv_hak=cond(status==1|status==2, ///
            (1995-yydx)*12+(12-mmdx), surv_mm)
```

```
.stset surv_hak,f(status==1 2) id(id) scale(12)
```

- Data are expanded by age and calendar period:

```
stsplit fu, at(0(1)20)
replace age = age + fu
gen int year = yydx + fu
```

- File popmort with reference rates is merged with patients data:

```
sort year sex age
merge year sex age using popmort, keep(rate) nokeep
```

Estimates with and without the np(#) option

- In this small data set the expected survival can be estimated both using np(#) option

```
.stexpect,ratevar(rate) at(0(1)20)) ///
                 out(apprmelanhak,replace)_np(100)
```

- and without using it

```
.stexpect,ratevar(rate) at(0(1)20) out(melanhak,replace)
```

These estimates are compared with the results produced by SURV3, a DOS program designed for the survival analysis based on cancer registry data.

Time Time	np(100)	SURV 3	Exact
1	0.97904	0.97904	0.97904
2	0.95808	0.95808	0.95808
3	0.93703	0.93704	0.93703
4	0.91594	0.91595	0.91594
5	0.89482	0.89483	0.89482
6	0.87361	0.87362	0.87361
7	0.85238	0.85238	0.85238
8	0.83113	0.83113	0.83113
9	0.80992	0.80992	0.80992
10	0.78880	0.78880	0.78880
11	0.76774	0.76773	0.76774
12	**0.74673**	**0.74656**	**0.74665**
13	**0.72553**	**0.72514**	**0.72536**
14	**0.70464**	**0.70399**	**0.70434**
15	**0.68410**	**0.68322**	**0.68366**
16	**0.66359**	**0.66259**	**0.66301**
17	**0.64362**	**0.64241**	**0.64290**
18	**0.62396**	**0.62246**	**0.62302**
19	**0.60424**	**0.60232**	**0.60308**
20	**0.58382**	**0.58147**	**0.58226**

Comparison of Results

Time	np(100)	SURV 3	Exact
12	0.74673	0.74656	0.74665
13	0.72553	0.72514	0.72536
14	0.70464	0.70399	0.70434
15	0.68410	0.68322	0.68366
16	0.66359	0.66259	0.66301
17	0.64362	0.64241	0.64290
18	0.62396	0.62246	0.62302
19	0.60424	0.60232	0.60308
20	0.58382	0.58147	0.58226

- The stexpect and SURV3 estimates differ from 12 years since diagnosis on, but always in a very small amount.
- Compared with “exact” results the np(#) approximation will be always biased upward. However in this example at the end of follow-up the bias is less than 0.002.

Observed Survival

To compute a ratio between observed and expected survival, the observed survival must be estimated at the same follow-up times specified when stexpect has been used:

```
use melanoma,clear
keep if year8594==0 & stage==1
stset surv_mm,f(status==1 2) id(id) scale(12)
stsplit fu, at(0(1)20)
sts gen Osservata = s Hilim = ub(s) Lowlim = lb(s)
```

Confidence intervals for log(-logS(t)) can be used to estimate confidence intervals for the Relative Survival.

Merging Estimates

Only one observation at the end of each follow-up time is kept:

```
bysort  _t : keep if _t==fu+1 & _n==1
```

After renaming _t, the file with observed estimates can be merged with the file with expected survival at the corresponding times:

```
keep _t Osservata Hilim Lowlim
rename _t t_exp
sort t_exp
merge t_exp using apprmelanhak
rename t_exp time
```

Relative Survival

```
gen double Relsurv = Osservata / Survexp
```

Confidence Intervals

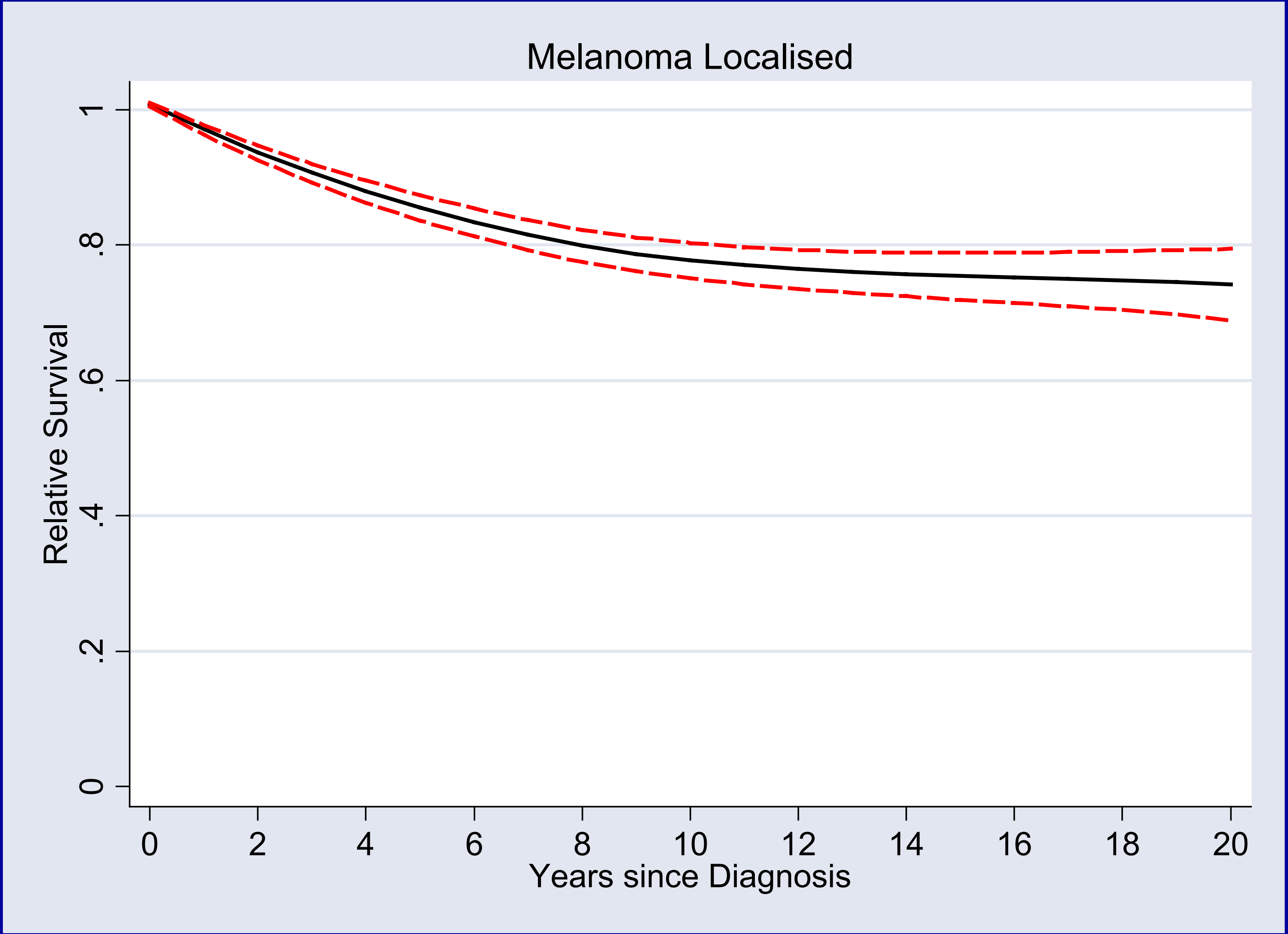
```
replace Hilim = Hilim / Survexp

replace Lowlim = Lowlim / Survexp
```

The results are tabulated sideways and graphed in the next slide.

```
+-------------------------------------------+
| time   Osservata     Survexp     Relsurv |
|-------------------------------------------|
|    0           1           1           1 |
|    1   .97062058   .97904406    .9913962 |
|    2   .90107875   .95807621    .9405084 |
|    3   .82824544    .9370337    .8839014 |
|    4   .78529246   .91594289    .8573596 |
|    5   .74700829   .89482277    .8348114 |
|-------------------------------------------|
|    6   .71057591    .8736134    .8133757 |
|    7   .67927507   .85237852    .7969171 |
|    8   .65496515   .83112553     .788046 |
|    9   .62972024   .80992114    .7775081 |
|   10    .6091457   .78880281    .7722408 |
|-------------------------------------------|
|   11   .58434944   .76773565    .7611337 |
|   12   .56019137   .74672549    .7501972 |
|   13   .54832636    .7255261    .7557638 |
|   14   .53309461   .70464163    .7565472 |
|   15   .51557421    .6840979    .7536556 |
|-------------------------------------------|
|   16   .49996838   .66358734    .7534326 |
|   17   .48209584   .64362049    .7490374 |
|   18   .47039178   .62396114      .75388 |
|   19   .45598877   .60424111     .754647 |
|   20   .42710142   .58381723     .731567 |
+-------------------------------------------+
```

```
woway  (lowess Relsurv time, clw(medthick) clc(black)) ///
       (lowess Hilim time, clc(red) clw(medthick) clp(dash)) ///
       (lowess Lowlim time, clw(medthick) clc(red) clp(dash)), ///
        xla(0(2)20) yla(0(.2)1) legend(off) t1t("Melanoma Localised") ///
       yti("Relative Survival") xti("Years since Diagnosis")
```



Period Analysis

- Period analysis is a relatively new method proposed by Brenner et al. to derive more up-to-date long-term relative survival estimates better describing the improvements in life expectancy of cancer patients.

- To obtain period survival estimates left truncated observations have to be allowed., i.e. subjects are allowed to enter the observation time after the diagnosis.

Is stexpect compatible with late entry?

- By the enter option in the stset command it is possible to deal with left truncation (late entry) in survival data.

- Internal codes of stexpect recognize the occurrence of late entry in the data and adapt its computations to this situation.

- Period estimates of relative survival can be achieved as illustrated previously.

strs

- strs is a new Stata command written by Paul Dickman and available at:

  http://www.pauldickman.com/rsmodel/stata_colon/

- This command estimates expected and relative survival according to the Conditional Method. The applied formula is somewhat different, assuming that data are grouped in time intervals.

- In my checks stexpect,method(2) and strs estimates are very similar.

Conclusions

- stexpect is a new "st" command. It takes advantage of all of the checks and flexibilities stset allows. Its use strictly depends on a timevar suitably generated by the user.

- It does not directly estimate relative survival, but few simple instructions are required to compute it.

- Estimates are consistent (at least until now) with the output of other programs. Only the spreading of stexpect may reveal its limits and contribute to its improvement.

References

- Therneau, T. E. and Grambsh, P. M. Modeling Survival Data –Extending the Cox Model. New York: Springer-Verlag (2000).
- Therneau T. and Offord J. Expected Survival Based on Hazard Rates (Update). Technical Report Number 63, Section of Biostatistic – Mayo Clinic.
- Voutilainen E. T., Dickmann P. W. and Hakulinen T. SURV2: Relative Survival Analysis Program – Software Manual http://www.cancerregistry.fi/surv2/
- Hakulinen T. Cancer survival corrected for heterogeneity in patient withdrawal. Biometrics, 38: 933–942, 1982.
- Brenner H, and Gefeller O. Deriving More Up-to-Date Estimates of Long-Term Patient Survival. J. Clin. Epidemiol., 50: 211-216, 1997.