

# Estimating relative survival using the `-strs-` command

Paul W. Dickman  
Department of Medical Epidemiology and Biostatistics  
Karolinska Institutet, Stockholm, Sweden  
paul.dickman@ki.se

6 September 2007

Workshop on methods for studying cancer patient survival with application in Stata, September 6, 2007

## Measuring the prognosis of cancer patients

- Total mortality (among the patients).
- Cause-specific mortality.
- Excess mortality.

$$\begin{array}{rcccl} \text{excess} & = & \text{total} & - & \text{expected} \\ \text{mortality} & & \text{mortality} & & \text{mortality} \end{array}$$

- Relative survival is the survival analog of excess mortality — the relative survival ratio is defined as the observed survival in the patient group divided by the expected survival of a comparable group from the general population.

Workshop on methods for studying cancer patient survival with application in Stata, September 6, 2007

1

## The colon carcinoma data set

```
-----  
sex          byte    sex      Sex  
age          byte    Age at diagnosis  
stage        byte    stage    Clinical stage at diagnosis  
mmdx         byte    Month of diagnosis  
yydx         int     Year of diagnosis  
surv_mm      float   Survival time in months  
surv_yy      float   Survival time in years  
status       byte    status   Vital status at last contact  
subsite      byte    colonsub Anatomical subsite of tumour  
year8594     byte    year8594 Year of diagnosis 1985-94  
agegrp       byte    agegrp   Age in 4 categories  
dx           int     Date of diagnosis  
exit         int     Date of exit  
id           float   Unique ID  
-----
```

Workshop on methods for studying cancer patient survival with application in Stata, September 6, 2007

2

## The population mortality file (popmort.dta)

```
. list
+-----+
| sex  _year  _age  prob  rate |
+-----+
1. | 1   1951    0  .96429  .0363632 |
2. | 1   1951    1  .99639  .0036165 |
3. | 1   1951    2  .99783  .0021724 |
4. | 1   1951    3  .99842  .0015812 |
5. | 1   1951    4  .99882  .0011807 |
+-----+
6. | 1   1951    5  .99893  .0010706 |
7. | 1   1951    6  .99913  .0008704 |
8. | 1   1951    7  .99905  .0009504 |
```

## The strs command for estimating and modelling relative survival using Stata

- Estimating relative survival.
  - cohort, period, or hybrid approach
  - choice of three methods for estimating expected survival (Ederer I, Ederer II, Hakulinen)
  - estimates can be standardised (by age for example)
  - saves estimates for subsequent modelling (or presentation in tables or graphs)
- Modelling excess mortality (relative survival)
  - several alternative approaches to estimating the model

## An example: localised colon carcinoma

```
. use colon if stage==1, clear
. stset surv_mm, fail(status==1 2) id(id) scale(12)
. strs using popmort, br(0 0.5 1(1)9) mergeby(_year sex _age) by(sex)
-> sex = Male
+-----+
| interval  n    d    w      p  p_star      r      cp  cp_e2  cr_e2 |
+-----+
| 0 .5  2620  229    0  0.9126  0.9728  0.9381  0.9126  0.9728  0.9381 |
| .5 1  2391  99    0  0.9586  0.9749  0.9833  0.8748  0.9484  0.9224 |
| 1 2  2292  229  166  0.8963  0.9483  0.9452  0.7841  0.8993  0.8719 |
| 2 3  1897  180  139  0.9015  0.9470  0.9519  0.7069  0.8517  0.8300 |
| 3 4  1578  140  119  0.9078  0.9449  0.9607  0.6417  0.8048  0.7974 |
+-----+
| 4 5  1319  113  104  0.9108  0.9428  0.9660  0.5845  0.7588  0.7703 |
| 5 6  1102  102   81  0.9039  0.9414  0.9601  0.5283  0.7143  0.7396 |
| 6 7   919   71   71  0.9196  0.9409  0.9774  0.4859  0.6721  0.7229 |
| 7 8   777   59   72  0.9204  0.9391  0.9800  0.4472  0.6312  0.7084 |
| 8 9   646   49   62  0.9203  0.9380  0.9811  0.4115  0.5921  0.6950 |
+-----+
```

## Syntax of the `strs` command

```
strs using filename [if exp] [in range]
[iweight=varname], breaks(numlist ascending)
mergeby(varlist) [by(varlist) ddiagage(varname)
diagyear(varname) attage(newvarname)
attyear(newvarname) survprob(varname) maxage(int 99)
standstrata(varname) brenner list(varlist)
potfu(varname) format(%fmt) ederer1 notables
level(int) save[(replace)] savind(filename[, replace])
savgroup(filename[, replace]) ]
```

the patient data file must be `stset` using the `id()` option with time since entry in years as the timescale before using `strs`

using `filename` specifies a file containing general population survival probabilities sorted by the variables specified in `mergeby()`.

## Life table quantities calculated by `strs`

<code>start</code>	Start of life table interval
<code>end</code>	End of life table interval
<code>n</code>	Number alive at start
<code>d</code>	Number of deaths during the interval
<code>d_star</code>	Expected number of deaths
<code>ns</code>	Number of survivors
<code>w</code>	Withdrawals (censorings) during the interval
<code>n_prime</code>	Effective number at risk
<code>y</code>	Person-time at risk
<code>p</code>	Interval-specific observed survival
<code>se_p</code>	Standard error of P
<code>lo_p</code>	Lower 95% CI for P
<code>hi_p</code>	Upper 95% CI for P
<code>p_star</code>	Interval-specific expected survival (Ederer II)
<code>r</code>	Interval-specific relative survival (Ederer II)
<code>se_r</code>	Standard error of R
<code>lo_r</code>	Lower 95% CI for R

<code>hi_r</code>	Upper 95% CI for R
<code>cp</code>	Cumulative observed survival
<code>se_cp</code>	Standard error of CP
<code>lo_cp</code>	Lower 95% CI for CP
<code>hi_cp</code>	Upper 95% CI for CP
<code>nu</code>	Estimated excess mortality rate, $(d-d\_star)/y$
<code>cp_e1</code>	Cumulative expected survival (Ederer I)
<code>cr_e1</code>	Cumulative relative survival (Ederer I)
<code>lo_cr_e1</code>	Lower 95% CI for CR (Ederer I)
<code>hi_cr_e1</code>	Upper 95% CI for CR (Ederer I)
<code>cp_e2</code>	Cumulative expected survival (Ederer II)
<code>cr_e2</code>	Cumulative relative survival (Ederer II)
<code>lo_cr_e2</code>	Lower 95% CI for CR (Ederer II)
<code>hi_cr_e2</code>	Upper 95% CI for CR (Ederer II)
<code>cp_hak</code>	Cumulative expected survival (Hakulinen)
<code>cr_hak</code>	Cumulative relative survival (Hakulinen)
<code>lo_cr_hak</code>	Lower 95% CI for CR (Hakulinen)
<code>hi_cr_hak</code>	Upper 95% CI for CR (Hakulinen)

## Estimates can be saved to a file

```
. use colon if stage==1, clear
. stset surv_mm, fail(status==1 2) id(id) scale(12)
. strsort using popmort, br(0(1)10) mergeby(_year sex _age) by(sex agegrp) save
. use grouped, clear
. gen n0=n[_n-4]
. list sex agegrp n0 cp cr_e2 lo_cr_e2 hi_cr_e2 if end==5, sepby(sex) noobs
+-----+
| sex agegrp n0 cp cr_e2 lo_cr_e2 hi_cr_e2 |
+-----+
| Male 0-44 161 0.7737 0.7881 0.7102 0.8486 |
| Male 45-59 462 0.7686 0.8233 0.7766 0.8636 |
| Male 60-74 1228 0.5945 0.7512 0.7128 0.7878 |
| Male 75+ 769 0.4131 0.7777 0.7067 0.8479 |
+-----+
| Female 0-44 136 0.7657 0.7709 0.6866 0.8358 |
| Female 45-59 531 0.7765 0.7953 0.7536 0.8314 |
| Female 60-74 1488 0.6993 0.7873 0.7588 0.8141 |
| Female 75+ 1499 0.4854 0.7816 0.7374 0.8249 |
+-----+
```

Workshop on methods for studying cancer patient survival with application in Stata, September 6, 2007

9

## Expected survival using three different methods

- To obtain estimates of expected survival using the Hakulinen method we must specify, using the `potfu()` option, a variable containing the last date of potential follow-up for each patient.
- The `ederer1` option results in Ederer I estimates of expected and relative survival also being estimated.
- Ederer II estimates are produced by default and no option is required.

```
. use colon, clear
. stset exit, origin(dx) fail(status==1 2) id(id) scale(365.24)
. gen long potfu = date("31/12/1995", "dmy")
. strsort using popmort if stage==1, br(0(1)10) mergeby(_year sex _age) \\\
by(sex) list(start n d w cr_e1 cr_e2 cr_hak) ederer1 potfu(potfu)
```

Workshop on methods for studying cancer patient survival with application in Stata, September 6, 2007

10

```
-> sex = Male
+-----+
| start end n d w cr_e1 cr_e2 cr_hak |
+-----+
| 0 1 2620 328 0 0.9238 0.9238 0.9238 |
| 1 2 2292 229 166 0.8758 0.8732 0.8756 |
| 2 3 1897 180 139 0.8361 0.8312 0.8359 |
| 3 4 1578 140 119 0.8050 0.7986 0.8049 |
| 4 5 1319 113 104 0.7787 0.7715 0.7787 |
+-----+
| 5 6 1102 102 81 0.7486 0.7407 0.7487 |
| 6 7 919 71 71 0.7333 0.7239 0.7335 |
| 7 8 777 59 72 0.7200 0.7095 0.7202 |
| 8 9 646 49 62 0.7082 0.6961 0.7082 |
| 9 10 535 33 58 0.7085 0.6948 0.7087 |
+-----+
```

Workshop on methods for studying cancer patient survival with application in Stata, September 6, 2007

11

## Estimating relative survival using a period approach

- Application of cohort, complete, period, or hybrid analysis is simply a matter of appropriately specifying the time at risk for each individual.
- For example, for period estimation with a 'window' between 1 January 1990 and 31 December 1994 (the last five years for which incidence data were collected in this dataset) we only consider person-time within this window.

```
. stset exit, origin(dx) enter(time mdy(1,1,1990)) ///  
  exit(time mdy(12,31,1994)) f(status==1 2) id(id) scale(365.24)
```

- We can then apply `strs` in the usual manner to obtain Ederer II estimates

```
. strs using popmort if stage==1, br(0(1)10) ///  
  mergeby(_year sex _age) by(sex)
```

or Hakulinen estimates

```
. replace potfu = date("31/12/1994","dmy")  
. strs using popmort if stage==1, br(0(1)10) potfu(potfu) ///  
  by(sex) mergeby(_year sex _age)
```

- See `colon_lifetable_strs_period.do`

## Modelling excess mortality (relative survival)

- The hazard at time since diagnosis  $t$  for persons diagnosed with cancer is modelled as the sum of the known baseline hazard,  $\lambda^*(t)$ , and the excess hazard due to a diagnosis of cancer,  $\nu(t)$  [1, 2, 3, 4, 5].

$$\lambda(t) = \lambda^*(t) + \nu(t)$$

- It is common to assume that the excess hazards are piecewise constant and proportional. Provides estimates of relative excess risk.
- Non-proportional excess hazards are common but can be incorporated by introducing follow-up time by covariate interaction terms.
- Can be estimated using a full-likelihood approach [2] or GLM with complementary log-log link [3].
- Poisson regression in a GLM framework is more practical and mathematically equivalent to the full-likelihood approach.

## Modelling excess mortality using Poisson regression

- The model can be written as

$$\ln(\mu_j - d_j^*) = \ln(y_j) + \mathbf{x}\beta, \quad (1)$$

where  $\mu_j = E(d_j)$ ,  $d_j^*$  the expected number of deaths, and  $y_j$  person-time.

- This implies a generalised linear model with outcome  $d_j$ , Poisson error structure, link  $\ln(\mu_j - d_j^*)$ , and offset  $\ln(y_j)$ .
- The user-defined link function is specified in an ado file (`rs.ado`).
- Such models have previously been described by Breslow and Day (1987) [6, pp. 173–176] and Berry (1983) [4].
- The usual regression diagnostics (residuals, influence statistics) and method for assessing model fit for generalised linear models can be utilised.

## First estimate relative survival and save the estimates COLON\_LIFETABLE\_STRS.DO

- Produces lifetable estimates of relative survival and stores the estimates in the data file `colon_grouped.dta`. The `notables` option suppresses printing of the lifetables in the output window.

```
use colon if stage==1, clear

stset surv_mm, fail(status==1 2) id(id) scale(12)

strs using popmort, br(0(1)10) mergeby(_year sex _age) ///
  by(sex year8594 agegrp) notables ///
  savind(colon_individ, replace) savgroup(colon_grouped, replace)
```

## Now fit the model to the first 5 years of follow-up COLON\_POISSON\_REGRESSION\_STEPFUNCTION.DO

```
. use colon_grouped if end < 6, clear
(Collapsed (or grouped) survival data)

. xi: glm d i.end i.sex i.year8594 i.agegrp , fam(pois) ///
  link(rs d_star) lnoffset(y) eform

Generalized linear models      No. of obs      =      80
Optimization      : ML      Residual df      =      70
Scale parameter =      1
Deviance      = 131.4342128      (1/df) Deviance = 1.877632
Pearson      = 130.1530694      (1/df) Pearson = 1.85933

Variance function: V(u) = u      [Poisson]
Link function      : g(u) = log(u-d*)      [Relative survival]

Log likelihood      = -245.9836017      AIC      = 6.39959
BIC      = -175.3077
```

	d	exp(b)	Std. Err.	z	P> z	[95% Conf. Interval]
_Iend_2		.7984084	.0730515	-2.46	0.014	.6673339 .955228
_Iend_3		.6230213	.0671961	-4.39	0.000	.5043086 .7696785
_Iend_4		.4969433	.0645561	-5.38	0.000	.3852391 .6410374
_Iend_5		.4334347	.065147	-5.56	0.000	.322838 .5819191
_Isex_2		.9564493	.0729823	-0.58	0.560	.8235891 1.110742
_Iyear8594		.7308044	.0539291	-4.25	0.000	.6323935 .8445296
_Iagegrp_1		.8642841	.1353083	-0.93	0.352	.635911 1.174672
_Iagegrp_2		1.071568	.1534869	0.48	0.629	.8092774 1.418869
_Iagegrp_3		1.436319	.2146593	2.42	0.015	1.071613 1.925147

### Fitted excess hazards



### Evidence of lack-of-fit!

- Residual deviance 131 on 70 df. Non-proportional excess hazards a possible culprit.
- Study of residuals (or subject-matter knowledge) suggests an age\*time interaction.

```
. xi: glm d i.end i.sex i.year8594 i.agegrp i.end*i.agegrp, ///
      fam(pois) link(rs_d_star) lnoffset(y) eform
```

```
Generalized linear models      No. of obs      =      80
Optimization      : ML      Residual df      =      58
Scale parameter =      1
Deviance      = 62.68719872    (1/df) Deviance = 1.080814
Pearson      = 59.38427551    (1/df) Pearson = 1.023867
```

```
Variance function: V(u) = u      [Poisson]
Link function      : g(u) = log(u-d*) [Relative survival]
```

```
Log likelihood = -211.6100946    AIC      = 5.840252
BIC      = -191.4703
```

- Interaction term highly statistically significant (LR test statistic 131-63=58 on 12df); no longer evidence of lack of fit.

- Estimated excess hazard ratios for the effect of age for each annual follow-up interval.

Age	Follow-up interval				
	1	2	3	4	5
0-44	1.00	1.00	1.00	1.00	1.00
45-59	1.10	0.59	1.22	0.72	0.97
60-74	1.66	0.89	1.02	0.85	0.97
75+	3.31	0.83	0.63	0.52	0.01

## References

- [1] Dickman PW, Sloggett A, Hills M, Hakulinen T. Regression models for relative survival. *Stat Med* 2004;**23**:51–64.
- [2] Estève J, Benhamou E, Croasdale M, Raymond L. Relative survival and the estimation of net survival: Elements for further discussion. *Statistics in Medicine* 1990;**9**:529–538.
- [3] Hakulinen T, Tenkanen L. Regression analysis of relative survival rates. *Applied Statistics* 1987;**36**:309–317.
- [4] Berry G. The analysis of mortality by the subject-years method. *Biometrics* 1983; **39**:173–184.
- [5] Pocock S, Gore S, Kerr G. Long term survival analysis: the curability of breast cancer. *Stat Med* 1982;**1**:93–104.
- [6] Breslow NE, Day NE. *Statistical Methods in Cancer Research: Volume II - The Design and Analysis of Cohort Studies*. IARC Scientific Publications No. 82. Lyon: IARC, 1987.