The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*), *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
Printed & electronic		**Printed & electronic**	
1-year subscription	$ 98	1-year subscription	$138
2-year subscription	$165	2-year subscription	$245
3-year subscription	$225	3-year subscription	$345
1-year student subscription	$ 75	1-year student subscription	$ 99
1-year institutional subscription	$245	1-year institutional subscription	$285
2-year institutional subscription	$445	2-year institutional subscription	$525
3-year institutional subscription	$645	3-year institutional subscription	$765
Electronic only		**Electronic only**	
1-year subscription	$ 75	1-year subscription	$ 75
2-year subscription	$125	2-year subscription	$125
3-year subscription	$165	3-year subscription	$165
1-year student subscription	$ 45	1-year student subscription	$ 45

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, Stata, Stata Press, Mata, Mata, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2014)
14, Number 1, pp. 176–190

strel2: A command for estimating excess hazard and relative survival in large population-based studies

Michael Hills
London School of Hygiene and Tropical Medicine, Retired
London, UK
mhills@blueyonder.co.uk

Bernard Rachet
London School of Hygiene and Tropical Medicine
London, UK
bernard.rachet@lshtm.ac.uk

Milena Falcaro
London School of Hygiene and Tropical Medicine
London, UK
milena.falcaro@lshtm.ac.uk

Abstract. In this article, we describe `strel2`, a Stata command for the estimation of excess hazard and relative survival in large population-based datasets. `strel2` implements the maximum-likelihood estimation approach developed by Estève et al. (1990, *Statistics in Medicine* 9: 529–538) and assumes that the excess hazard is a piecewise constant function. Categorical covariates can be incorporated into the model, allowing the user to obtain interval- and covariate-specific estimates of the quantities of interest. Although alternative and more powerful commands for relative survival exist, `strel2` is a simple tool that is particularly convenient for users who may not have strong statistical skills and want to analyze very large datasets.

Keywords: st0330, strel2, background mortality, excess hazard, relative survival, net survival

1 Introduction

When the prognosis of a given disease is investigated, clinicians, public health scientists, health policy makers, and the like are generally interested in the mortality due to the disease of interest rather than in the overall mortality. This notion (the net mortality) and its complement (the net survival), first introduced more than 60 years ago by Berkson and Gage (1950), are defined as the mortality or survival the patients would experience if they could only die from the disease of interest. In the context of population-based cancer data, relative survival was proposed in the late fifties as an approach to estimate the net survival (for example, Ederer and Heise [1959] and Ederer, Axtell, and Cutler [1961]). Relative survival can be defined as the survival of

the patients after removing the "background" or general population mortality, that is, the survival that would be expected if patients had the same mortality as the general population from which they came. Several methods have been proposed for the estimation of net survival. These methods include the nonparametric estimator by Pohar Perme, Stare, and Estève (2012) and the parametric models by Estève et al. (1990), Dickman et al. (2004), and Nelson et al. (2007).

Here we describe `strel2`, a Stata command based on the maximum-likelihood estimation approach developed in Estève et al. (1990). It was first implemented in the late nineties under the name of `strel` to estimate relative survival of cancer patients diagnosed in England and Wales (Coleman et al. 1999). `strel` was especially designed to be applied on massive datasets, but it did not allow covariates. The new `strel2` command enables estimation of the effects of covariates (excess-hazard ratios) as well as their associated relative survival.

2 The maximum likelihood approach

Estève et al. (1990) proposed a full maximum likelihood approach for the estimation of excess hazard and net survival from individual records. In their model, time since diagnosis is divided into intervals, and λ_{ij}, the observed hazard of death for subject i in interval j, is assumed to be the sum of two components: the background or general population hazard λ^*_{ij} and the cancer-related or excess hazard ν_{ij}; that is,

$$\lambda_{ij} = \lambda^*_{ij} + \nu_{ij}$$

The term λ^*_{ij} is the background hazard for subject i at the age reached in interval j and so depends on the age at diagnosis of subject i and the interval. It is a known function that is usually retrieved from life tables and typically varies at least by sex, calendar period or year, and geographical area. In the absence of covariates, the excess hazard ν_{ij} is constant for all subjects during time interval j, and we shall write it as α_j. In the presence of a covariate, ν_{ij} is modeled as $\alpha_j e^{\beta x_i}$, where x is the covariate and β is a parameter. The contribution to the total log likelihood from each subject interval is

$$-\lambda_{ij} y_{ij} + d_{ij} \ln(\lambda_{ij})$$

where d_{ij} is the corresponding event indicator (1 if the subject died in that interval and 0 otherwise) and y_{ij} is the follow-up time in interval j. Substituting for λ_{ij} shows that, apart from a constant, this is equal to

$$-\nu_{ij} y_{ij} + d_{ij} \ln(\lambda^*_{ij} + \nu_{ij})$$

This log likelihood is summed within strata formed by the values of any covariates. For simplicity, we shall consider the case where there are no covariates, that is, $\nu_{ij} = \alpha_j$. The first term is contributed by every subject interval, and adding over subjects, we get $-\alpha_j Y_j$ for interval j, where $Y_j = \sum_i y_{ij}$. The second term reduces to 0 unless there is a death in interval j, in which case it becomes $\ln(\lambda^*_{ij} + \alpha_j)$. The total log likelihood is, therefore,

$$-\sum_j \alpha_j Y_j + \sum_{\text{deaths}} \ln(\lambda^*_{ij} + \alpha_j)$$

where $\sum_{\text{deaths}}$ denotes a sum over i and j such that $d_{ij} = 1$.

3 The strel2 command

`strel2` uses the same model as Estève et al. (1990), but it estimates excess hazard and relative survival using a variant of the maximum likelihood approach they used. Survival times are stratified by interval and covariate values, and person-years (Y_j) are first calculated for each of these strata. The stratum-specific person-years are then divided equally between the stratum's event records (that is, those that end in a failure). In this way, the contribution to the likelihood from each stratum can be computed using only the information stored in the event records, and no data splitting is required. Of course, this approach requires observing at least one event per stratum and, therefore, is not convenient or applicable to small or sparse datasets. It becomes particularly advantageous with large cohort studies, where data splitting would have severe repercussions on the computational time; however, this advantage in computational time is progressively lost as more covariate values are added because of the need to stratify by these values.

This simple command has its drawbacks; namely, continuous variables and time-varying effects cannot be included. The latter means that the model assumes proportionality of the excess hazards. Nonetheless, combined with its speed, the simplicity of `strel2` fits the needs of many cancer registries with large datasets.

The log likelihood is maximized using the Stata command `ml lf`, which requires the user to supply a subroutine to calculate the log likelihood for each observation. In this case, the subroutine is named `strel2_llc.ado` and is included in the package.

3.1 Syntax

`strel2` [*xvars*] `using` *filename* [*if*] [*in*] `, breaks(`*numlist*`) mergeby(`*varlist*`)`
[`level(`#`) negalpha iter(`#`) eform diagdate(`*varname*`) period(`*numlist*`)`
`hybrid trace group rtable basetable display saving(`*string*`) replace`]

where *filename* is the file containing the general population mortality rates, and *xvars* is a list of covariates, each of which must be a categorical variable with not more than 20 categories. String covariates must be numerically coded, and this is most easily done using the Stata command `encode`. For example, to encode a string variable `Z` and place the numeric codes in the numeric variable `Znum`, use `encode Z, generate(Znum)`. The numeric codes are labeled with the original string values. The command automatically treats covariates as factors, so the prefix `i.` should not be used.

The main patient data cannot include variables named `age` (attained age), `rate` (general population mortality), or `_interval`, while the life table file must contain `age`, `rate`, and all the variables specified in `mergeby()`. Before using `strel2`, data must be `stset`. More specifically, the `stset` command must include either the `enter()` option or the `origin()` option, or both of them; that is, the survival times cannot be calculated a priori by the user and then declared as survival data with `stset` without specifying the age at entry or the origin or both and the age at exit. Records with missing values on the survival time, the event indicator, or any of the *xvars* are disregarded.

3.2 Options

`breaks(`*numlist*`)` specifies the cutpoints defining intervals over which the baseline excess hazard is assumed to be constant. For example, `breaks(0 1 2.5 10)` implies intervals $(0, 1)$, $(1, 2.5)$, and $(2.5, 10)$. The first element in *numlist* is usually 0; if this is not the case, the estimates provided by `strel2` are conditional upon surviving to the time point specified by this first element. The last intervals included in the analysis with no deaths and no person-years are simply dropped from the analysis. `breaks()` is required.

`mergeby(`*varlist*`)` identifies the variables that, along with the variable `age` (attained age), will be used to merge the patient data with the general population mortality rates. The order of the variables in *varlist* is not important, but `age` and *varlist* must be present in the life table file and must uniquely identify its records. On the other hand, the main dataset must contain all the variables in *varlist* but no variable named `age`. Before the merging, `strel2` internally creates a new variable called `age` in the patient data file. The general population mortality rates are then merged using `age` and *varlist* as matching variables. `mergeby()` is required.

`level(#)` specifies the confidence level, as a percentage, for confidence intervals (CIs). The default is `level(95)` or as set by `set level`.

`negalpha` causes negative interval-specific excess hazards to be replaced with 0. In some circumstances, the hazard of death experienced by the cancer patients becomes lower than the mortality hazard of the general population to which the cancer patients are compared. It may occur with sparse data, but it may also reflect data issues such as inadequacy of the life tables or high proportion of the so-called "immortals", that is, patients who are lost to passive follow-up. This option is rarely advisable and should be used with great caution.

`iter(#)` sets the maximum number of iterations. The default is `iter(100)`.

`eform` reports the exponentiated coefficients and corresponding standard errors and CIs for the covariates. The coefficients for the time intervals are never exponentiated.

`diagdate(`*varname*`)` indicates the name of the variable containing the date at diagnosis in the patient dataset.

`period(`*numlist*`)` requests the use of the period approach (Brenner and Gefeller 1996). Here *numlist* is a list of integers specifying the year(s) of diagnosis for which survival will be predicted. This approach enables survival prediction for recently diagnosed cancer patients. For example, we can estimate 5-year relative survival for patients diagnosed in the year 2000 and followed up to the end of the year 2000 by using conditional survival probabilities of patients diagnosed in the past and alive at a certain point in the year 2000. In this case, we would specify the `period(2000)` option. If an interval rather than a single year is considered, the upper and lower boundaries of that interval are specified as two 4-digit calendar years separated by a space. The option `period(2000 2002)`, for instance, means that only patients alive at a certain time between the years 2000 and 2002 will contribute to the survival estimation. When `period()` is used, the survival times must have been `stset` in years using dates (see comments on `stset` in section 4), and the `diagdate(`*varname*`)` option must be specified as well.

`hybrid` can be used in conjunction with `period()` and `diagdate()` to request the hybrid approach proposed by Brenner and Rachet (2004). This is an extension of the period approach that is used when the reporting of new incident cases lags behind their follow-up: for example, incidence is complete up to the year 2000, but deaths up to the year 2002 are known and included in the data. The hybrid approach avoids the bias that may arise in this context with the period approach by ensuring that the same number of conditional survival probabilities is available for each of the follow-up years. For instance, by specifying `period(2001 2002) diagdate(datediag) hybrid`, the survival is estimated for patients who were diagnosed with cancer in the years 2001–2002.

`trace` displays the number of deaths and person-years for each interval and combination of covariate values.

`group` redefines the breakpoints when there is at least one stratum with person-years but zero deaths. More specifically, it combines intervals by removing the cutpoints that are right-hand ends of intervals with no event and nonzero person-years. For example, let's consider the following situation where we use `breaks(0 1 2 3 5 7)`:

```
breaks          0     1     2      3          5          7
                |-----|-----|------|----------|----------|
deaths             3     0      5         9          6
person-years     200   140    100       150         90
```

Without the `group` option, `strel2` would not perform the estimation because in the second interval there are 140 person-years but no deaths. However, if we use `group` in the `strel2` command line, the model would be fit using 0, 1, 3, 5, and 7 as the new cutpoints:

```
new breaks      0     1            3          5          7
                |-----|------------|----------|----------|
deaths             3          5          9          6
person-years     200        240        150         90
```

`rtable` displays the interval-specific observed death rates and survival probabilities for the cohort under study.

`basetable` displays the estimated individual-specific excess hazards and relative survivals for the reference category. This is set as the default if the model has no covariate.

`display` is an option that has an effect only when the user includes covariates in the model. It forces Stata to show on the screen the interval-specific point and interval estimates of the excess hazard and relative survival for each combination of covariate values.

`saving(`*string*`)` saves the results in a Stata data file. The relative survival estimates are stored in a variable called `RelS`, whereas the estimated excess hazards are saved in `alpha`. The lower and upper bounds of the CIs are denoted with suffixes `_lo` and `_up`, respectively. Intervals are defined with the `start` and `end` variables. If covariates were included in the model, the results file also contains them because the estimates are interval and covariate specific. However, if the covariate variable names overlap with the name of any of the variables used for storing the results, their names will be prefixed by `c_`. For example, a covariate `alpha` would be saved as `c_alpha`.

`replace` allows Stata to overwrite the file specified in `saving()` if it already exists.

4 Examples

To illustrate the use of the `strel2` command, we consider a subsample of 90,000 anonymized female patients who were diagnosed with colorectal cancer in England between the years 1996 and 2006 with follow-up to the end of the year 2009. The values of the date variables were randomly modified to further enhance data privacy and prevent disclosure of individual information. A description of the variables relevant to the analysis is reported in table 1.

Table 1. Description of the variables relevant to the analysis

Variable name	Description	Values
`dead`	event indicator	`0` = observation is right-censored and `1` = event (death)
`agediag`	age at diagnosis in years	min = `15.1` and max = `100`
`ageout`	age at exit	min = `18.3` and max = `108.5`
`agegrp`	categorized age at diagnosis	`15` = $[15, 45)$, `45` = $[45, 55)$, `55` = $[55, 65)$, `65` = $[65, 75)$, `75` = $[75, 85)$, `85` = $85+$
`_year`	year at exit of the study	min = `1996` and max = `2009`
`sex`	gender	`1` = male and `2` = female
`dep`	deprivation quintile	five categories where `1` = least deprived and `5` = most deprived
`gor`	government office region	nine categories
`registry`	cancer registry identifier	`1` to `8` (for example, `7` = West Midlands registry)

Before using `strel2`, we have to `stset` the data:

```
stset ageout, fail(dead) enter(time agediag)
```

Alternatively, if our dataset contains information on the date of diagnosis (`datediag`), the date of exit (`dateout`), and the date of birth (`dob`), we can use

```
stset dateout, fail(dead) origin(time dob) enter(time datediag) scale(365.25)
```

This latter is the `stset` format to be used before an `strel2` command with the `period()` option.

As mentioned earlier, when using `strel2`, the survival times (`survt` = `ageout` - `agediag`) cannot be directly declared as survival data with `stset survt, fail(dead)`. Because the time scale is age, either the `origin()` option or the `enter()` option or both must be used.

Hereafter, we will use the general population mortality rates that are stored in `lifetable.dta` and are stratified by `age`, `sex`, `dep`, `_year`, and `gor`.

A simple model

We first describe an `strel2` model without covariates. We expect the excess-hazard function to change quite rapidly early in the follow-up, so we consider monthly intervals at the beginning and longer intervals thereafter.

```
. strel2 using lifetable, breaks(0(.08333333)0.5 0.75(0.25)2 3(1)8 10)
> mergeby(sex dep _year gor) rtable
Breaks  0 .08333333 .16666666 .24999999 .33333332 .41666665 .49999998 .75 1
> 1.25 1.5 1.75 2 3 4 5 6 7 8 10
```

(*output omitted*)

```
Estimation of alpha                               Number of obs   =      60392
                                                  Wald chi2(18)   =   28471.17
Log likelihood = -112225.65                       Prob > chi2     =     0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_interval						
2	-1.075549	.0197123	-54.56	0.000	-1.114184	-1.036914
3	-1.289666	.0189058	-68.22	0.000	-1.32672	-1.252611
4	-1.418926	.0183666	-77.26	0.000	-1.454924	-1.382928
5	-1.463175	.0181833	-80.47	0.000	-1.498814	-1.427537
6	-1.503049	.0180254	-83.39	0.000	-1.538378	-1.46772
7	-1.535798	.0169289	-90.72	0.000	-1.568978	-1.502618
8	-1.557938	.0169084	-92.14	0.000	-1.591078	-1.524798
9	-1.584892	.0168739	-93.93	0.000	-1.617964	-1.55182
10	-1.604488	.01685	-95.22	0.000	-1.637513	-1.571462
11	-1.624509	.0168187	-96.59	0.000	-1.657473	-1.591545
12	-1.643647	.0167841	-97.93	0.000	-1.676544	-1.610751
13	-1.672735	.0164982	-101.39	0.000	-1.705071	-1.640399
14	-1.704103	.0164846	-103.38	0.000	-1.736413	-1.671794
15	-1.720215	.0164826	-104.37	0.000	-1.75252	-1.687909
16	-1.735237	.0164751	-105.33	0.000	-1.767527	-1.702946
17	-1.737528	.0164844	-105.40	0.000	-1.769837	-1.70522
18	-1.744477	.0164877	-105.80	0.000	-1.776792	-1.712161
19	-1.751873	.0164544	-106.47	0.000	-1.784123	-1.719623
_cons	1.75823	.0164187	107.09	0.000	1.72605	1.79041

Observed death rates and survival probabilities

start	end	subj	dths	pyrs	d_rate	Surv	lower	upper
0.00	0.08	90000	12489	6802.51	1.8359	85.81	85.58	86.04
0.08	0.17	77511	4664	6243.75	0.7470	80.63	80.37	80.89
0.17	0.25	72847	3121	5931.11	0.5262	77.17	76.90	77.45
0.25	0.33	69725	2246	5712.85	0.3931	74.69	74.40	74.97
0.33	0.42	67478	1909	5542.42	0.3444	72.57	72.28	72.87
0.42	0.50	65567	1649	5395.00	0.3057	70.75	70.45	71.05
0.50	0.75	63916	4170	15441.20	0.2701	66.13	65.82	66.44
0.75	1.00	59743	3525	14484.67	0.2434	62.23	61.91	62.54
1.00	1.25	56217	2938	13685.52	0.2147	58.98	58.65	59.30
1.25	1.50	53276	2524	12994.61	0.1942	56.18	55.85	56.50
1.50	1.75	50744	2148	12411.71	0.1731	53.80	53.47	54.13
1.75	2.00	48591	1824	11910.72	0.1531	51.78	51.45	52.11
2.00	3.00	46763	5501	43822.00	0.1255	45.67	45.35	46.00
3.00	4.00	41246	3539	37379.13	0.0947	41.55	41.22	41.87
4.00	5.00	33894	2515	30712.75	0.0819	38.28	37.96	38.60
5.00	6.00	27742	1727	25209.83	0.0685	35.74	35.42	36.07
6.00	7.00	22745	1406	20490.69	0.0686	33.37	33.05	33.70
7.00	8.00	18361	1086	16437.31	0.0661	31.24	30.91	31.57
8.00	10.00	14578	1411	22532.56	0.0626	27.56	27.22	27.90

Excess mortality and relative survival probabilities

start	end	alpha	alpha_lo	alpha_up	RelS	RelS_lo	RelS_up
0.00	0.08	1.7582	1.72605	1.79041	86.37	86.14	86.60
0.08	0.17	0.6827	.6613004	.7040617	81.59	81.33	81.86
0.17	0.25	0.4686	.4501935	.4869353	78.47	78.19	78.75
0.25	0.33	0.3393	.323171	.3554377	76.28	75.99	76.57
0.33	0.42	0.2951	.2797398	.3103695	74.43	74.13	74.73
0.42	0.50	0.2552	.2406005	.2697613	72.86	72.56	73.17
0.50	0.75	0.2224	.2143475	.2305164	68.92	68.60	69.24
0.75	1.00	0.2003	.1923744	.2082101	65.56	65.22	65.89
1.00	1.25	0.1733	.1657089	.1809676	62.78	62.44	63.11
1.25	1.50	0.1537	.1463179	.1611664	60.41	60.06	60.75
1.50	1.75	0.1337	.1265752	.1408671	58.42	58.07	58.77
1.75	2.00	0.1146	.1077559	.1214094	56.77	56.42	57.13
2.00	3.00	0.0855	.0823256	.0886643	52.12	51.76	52.48
3.00	4.00	0.0541	.0512426	.0570107	49.37	49.00	49.75
4.00	5.00	0.0380	.0351734	.0408573	47.53	47.15	47.92
5.00	6.00	0.0230	.0203256	.025661	46.45	46.06	46.85
6.00	7.00	0.0207	.0178202	.023583	45.50	45.09	45.91
7.00	8.00	0.0138	.0108012	.0167054	44.88	44.45	45.30
8.00	10.00	0.0064	.0042358	.0084775	44.31	43.85	44.77

The first output table shows the estimation results. In the `break()` option, we specified 20 cutpoints, so the baseline excess-hazard function is approximated with a 19-interval piecewise constant hazard function that in `strel2` is estimated using the 19-category variable `_interval`. The coefficient of `_cons` (that is, 1.75823) represents the excess hazard for the reference category, that is, interval 1. The estimated excess hazards for intervals 2 to 19 are, respectively, 0.682681 (that is, 1.75823 − 1.075549), 0.468564 (that is, 1.75823 − 1.289666), and so on. It is also worth noting that, as we mentioned earlier, `strel2` carried out the estimation by using only the data records where `_d` = 1. The number of observations reported above the table of estimated coefficients corresponds, therefore, to the number of deaths (60,392) observed during the follow-up under study.

Because we specified the `rtable` option, the output also includes a table with additional interval-specific information. More specifically, for each interval, it reports the number of subjects (`subj`) and deaths (`dths`), the person-years (`pyrs`), the observed death rate (`d_rate` = `dths`/`pyrs`), the observed survival (`Surv`), and its CIs (`lower` and `upper`).

The last table in the above output shows the interval-specific estimates of the excess hazard (`alpha`) and of the relative survival (`RelS`). Suffixes `_lo` and `_up` are used to denote, respectively, the lower and upper bounds of the CIs.

Including covariates in the model

Now let's suppose we are interested in estimating the effect of age at diagnosis. Because `strel2` allows only the inclusion of categorical covariates, we will use `agegrp` (that is, the categorized age at diagnosis) instead of `agediag`.

```
. strel2 agegrp using lifetable, breaks(0(.08333333)0.5 0.75(0.25)2 3(1)8 10)
> mergeby(sex dep _year gor)
```

(*output omitted*)

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
eq1						
agegrp						
45	.0900292	.0384053	2.34	0.019	.0147562	.1653023
55	.1157211	.0352128	3.29	0.001	.0467053	.184737
65	.2397361	.0339204	7.07	0.000	.1732534	.3062188
75	.5218398	.0335244	15.57	0.000	.4561331	.5875465
85	1.073545	.0342834	31.31	0.000	1.006351	1.14074
eq2						
_interval						
2	-.664259	.0247456	-26.84	0.000	-.7127595	-.6157585
3	-.7988362	.0283868	-28.14	0.000	-.8544734	-.7431991
4	-.8809898	.0306846	-28.71	0.000	-.9411306	-.820849
5	-.9100639	.0315135	-28.88	0.000	-.9718292	-.8482987
6	-.9349265	.0322286	-29.01	0.000	-.9980934	-.8717596
7	-.9559756	.0326175	-29.31	0.000	-1.019905	-.8920465
8	-.9706401	.0330616	-29.36	0.000	-1.03544	-.9058407
9	-.988102	.0335883	-29.42	0.000	-1.053934	-.9222702
10	-1.001074	.0339803	-29.46	0.000	-1.067674	-.9344739
11	-1.014609	.034387	-29.51	0.000	-1.082007	-.9472122
12	-1.02743	.0347725	-29.55	0.000	-1.095583	-.9592771
13	-1.04652	.0352969	-29.65	0.000	-1.115701	-.9773397
14	-1.067735	.0359467	-29.70	0.000	-1.138189	-.9972808
15	-1.078621	.0362833	-29.73	0.000	-1.149735	-1.007507
16	-1.089319	.0366098	-29.75	0.000	-1.161073	-1.017566
17	-1.090625	.0366526	-29.76	0.000	-1.162463	-1.018787
18	-1.095429	.0368032	-29.76	0.000	-1.167562	-1.023296
19	-1.101099	.0369653	-29.79	0.000	-1.17355	-1.028649
_cons	1.106415	.0371061	29.82	0.000	1.033689	1.179142

The estimation results are now divided into two parts or equations. The bottom part (labeled `eq2`) refers to the interval-specific excess-hazard indicators and is similar to the output we obtain when no covariate is included in the model. The first part (`eq1`) shows the effect of the age groups on the log excess-hazard scale. Here the `agegrp` variable is treated as a factor variable. `agegrp` was created with `egen agegrp=cut(agediag), at(15 45 55 65 75 85 100)`, so its values correspond to the left-hand bounds of the grouping intervals (that is, 15, 45, 55, 65, 75, and 85). The values below `agegrp` in `eq1` correspond to, respectively, `45.agegrp` (that is, the indicator variable for `agegrp` = 45), `55.agegrp`, `65.agegrp`, etc. By using the `eform` option, we can get the exponentiated form of the coefficients, that is, the excess-hazard ratios associated with the various age groups. For example, the age group 75–84 shows a log excess-hazard ratio of 0.5218398, which corresponds to an excess hazard ratio of 1.69 (95% CI: [1.58, 1.80]). It means that the patients diagnosed at age 75–84 experience a 69% higher risk of death from their cancer than the patients diagnosed between the ages of 15 and 44. It is worth noting that the excess hazards are proportional, that is, constant along time since diagnosis.

Had we specified the option `basetable` in the previous command, we would have also obtained additional output showing the point and interval estimates of the excess hazard and relative survival for the reference category, which in our analysis corresponds to patients whose age at diagnosis is between 15 and 44 years (that is, `agegrp` = 15).

By using the `display` option, we can request that the interval-specific estimates be computed and displayed on the screen for each combination of covariate values. The user can, therefore, easily obtain estimates of the excess hazard and relative survival for subgroups of interest. For simplicity, if we consider a model with only three intervals, and we specify the `display` and `eform` options, the output would be the following:

```
. strel2 agegrp using lifetable, breaks(0 1 5 10) mergeby(sex dep _year gor)
> eform display
```

(*output omitted*)

	exp(b)	Std. Err.	z	P>|z|	[95% Conf.	Interval]
eq1						
agegrp						
45	1.098995	.0421709	2.46	0.014	1.019373	1.184836
55	1.135696	.0399486	3.62	0.000	1.060036	1.216756
65	1.303832	.0441901	7.83	0.000	1.220035	1.393385
75	1.768615	.0592707	17.01	0.000	1.65618	1.888684
85	3.25117	.1115836	34.35	0.000	3.039664	3.477393
eq2						
_interval						
2	-.2289501	.0075849	-30.18	0.000	-.2438163	-.2140839
3	-.2760976	.0090585	-30.48	0.000	-.293852	-.2583431
_cons	.2885554	.0094208	30.63	0.000	.2700909	.3070198

-> agegrp = 15

start	end	alpha	alpha_lo	alpha_up	RelS	Re_lo	Re_up
0	1	.2885554	.2700909	.3070198	74.93453	73.51922	76.28687
1	5	.0596053	.0557281	.0634824	59.03874	57.59987	60.4468
5	10	.0124578	.011204	.0137116	55.47346	54.07952	56.84341

-> agegrp = 45

start	end	alpha	alpha_lo	alpha_up	RelS	Re_lo	Re_up
0	1	.3171208	.3040591	.3301825	72.82428	71.85966	73.76215
1	5	.0655059	.0626927	.068319	56.03777	55.06567	56.99778
5	10	.0136911	.0124828	.0148994	52.33004	51.36938	53.2812

```
-> agegrp = 55
```

start	end	alpha	alpha_lo	alpha_up	RelS	Re_lo	Re_up
0	1	.3277112	.318538	.3368844	72.05711	71.38988	72.7119
1	5	.0676935	.0656143	.0697726	54.96437	54.281	55.64209
5	10	.0141483	.0129625	.0153341	51.21046	50.50507	51.91108

(*output omitted*)

For example, the estimated excess hazard in the second interval (that is, 1–5) for a patient with `agegrp` = 55 is equal to `alpha` $= (0.2885554 - 0.2289501) \times 1.135696 = 0.0676935$, whereas the corresponding estimated relative survival, as a percentage, at 5 years is `RelS` $= 100 \times \exp\{-(0.3277112 \times 1 + 0.0676935 \times 4)\} = 54.964$.

Specifying the group option

If we now restrict our attention to the West Midlands cancer registry, that is, we specify `if registry==7` in the command line, we get the error message `1 stratum observed with p-years but no deaths`. By specifying the `trace` option, we can easily see that for the first age group, no death was observed in interval 7–8. If we add the `group` option, `strel2` refits the model after widening the time intervals with person-years > 0 but no deaths. In the following example, the `group` option leads to the merging of intervals $(7, 8)$ and $(8, 10)$ into a single time interval:

```
. strel2 agegrp using lifetable if registry==7,
> breaks(0(.08333333)0.5 0.75(0.25)2 3(1)8 10) mergeby(sex dep _year gor)
> group basetable

Breaks  0 .08333333 .16666666 .24999999 .33333332 .41666665 .49999998 .75 1
> 1.25 1.5 1.75 2 3 4 5 6 7 8 10
WARNING: 1 stratum observed with p-years but no deaths

Regrouping ...
New breaks  0 .08333333 .16666666 .24999999 .33333332 .41666665 .49999998 .75 1
> 1.25 1.5 1.75 2 3 4 5 6 7 10
```

(*output omitted*)

Excess mortality and relative survival probabilities for reference category

start	end	alpha	alpha_lo	alpha_up	RelS	RelS_lo	RelS_up
0.00	0.08	1.3423	1.068092	1.616534	89.42	87.18	91.28
0.08	0.17	0.5170	.403872	.6301225	85.65	83.38	87.63
0.17	0.25	0.3262	.2499147	.402429	83.35	81.10	85.36
0.25	0.33	0.2916	.2217234	.3615423	81.35	79.11	83.37
0.33	0.42	0.1991	.1471526	.2510572	80.01	77.79	82.04
0.42	0.50	0.1881	.1378792	.2383691	78.77	76.56	80.79
0.50	0.75	0.1934	.150457	.2363493	75.05	72.80	77.14
0.75	1.00	0.1574	.1211558	.193649	72.15	69.90	74.27
1.00	1.25	0.1419	.1083194	.1755591	69.64	67.38	71.77
1.25	1.50	0.1211	.09125	.1508777	67.56	65.32	69.69
1.50	1.75	0.1046	.077699	.1314702	65.82	63.59	67.94
1.75	2.00	0.1005	.0740684	.1268481	64.18	61.98	66.30
2.00	3.00	0.0711	.0550961	.0870664	59.78	57.51	61.97
3.00	4.00	0.0440	.0327763	.0551656	57.21	54.95	59.40
4.00	5.00	0.0306	.0213965	.0398436	55.48	53.23	57.67
5.00	6.00	0.0123	.005777	.0188362	54.81	52.55	57.00
6.00	7.00	0.0175	.0090939	.0259337	53.85	51.60	56.05
7.00	10.00	0.0106	.005345	.0158861	52.17	49.83	54.45

5 Conclusions

In this article, we have presented `strel2`, a Stata command that estimates excess mortality hazard and relative survival, and that combines simplicity and speed. Population-based cancer survival is a major indicator for public health and health care policy. The estimation of such an indicator, particularly in the context of international studies, is often based on large datasets with a very limited number of covariables. Age standardization is often required when comparing survival across different subpopulations because cancer survival often depends on age, and the age distribution of cancer patients varies between populations. After using `strel2`, age-standardized survival and its standard error can be obtained in the conventional way as described in Corazziari, Quinn, and Capocaccia (2004).

Other Stata commands for multivariable excess-hazard regression models exist. For example, `stpm2` (Lambert and Royston 2009) and `strs` (Dickman, Coviello, and Hills 2007) have several advantages over `strel2`. Nevertheless, `strel2` is a simple tool that is particularly convenient for very large datasets and for users who may not have strong statistical skills. It is important to highlight that the estimates from `strel2` are based on individual-level data and not on grouped data as in `strs`. The main limitation of `strel2` is its inability to relax the assumption of proportional excess hazards; that is, it is not possible to incorporate time-dependent interactions in the model. The full-likelihood approach models described in this article can also be fit by expanding the data with `strs` and then using a generalized linear model. This would allow the user to include continuous covariates and to account for nonproportional excess hazards; however, with large datasets, this could be computationally intensive if not prohibitive.

6 Acknowledgments

This work was supported by Cancer Research UK. The authors would like to thank Bianca De Stavola, Andy Sloggett, Adrian Mander, and Tony Brady for their help in the development of earlier versions of `strel2`.

7 References

Berkson, J., and R. P. Gage. 1950. Calculation of survival rates for cancer. *Mayo Clinic Proceedings* 25: 270–286.

Brenner, H., and O. Gefeller. 1996. An alternative approach to monitoring cancer patient survival. *Cancer* 78: 2004–2010.

Brenner, H., and B. Rachet. 2004. Hybrid analysis for up-to-date long-term survival rates in cancer registries with delayed recording of incident cases. *European Journal of Cancer* 40: 2494–2501.

Coleman, M. P., P. Babb, P. Damiecki, P. Grosclaude, S. Honjo, J. Jones, G. Knerer, A. Pitard, M. Quinn, A. Sloggett, and B. De Stavola. 1999. *Cancer Survival Trends in England and Wales, 1971–1995: Deprivation and NHS Region*. London: Stationery Office.

Corazziari, I., M. Quinn, and R. Capocaccia. 2004. Standard cancer patient population for age standardising survival ratios. *European Journal of Cancer* 40: 2307–2316.

Dickman, P. W., E. Coviello, and M. Hills. 2007. Estimating and modelling relative survival. http://www.pauldickman.com/survival/strs.pdf.

Dickman, P. W., A. Sloggett, M. Hills, and T. Hakulinen. 2004. Regression models for relative survival. *Statistics in Medicine* 23: 51–64.

Ederer, F., L. M. Axtell, and S. J. Cutler. 1961. The relative survival rate: A statistical methodology. *National Cancer Institute Monograph* 6: 101–121.

Ederer, F., and H. Heise. 1959. Instructions to IBM 650 programmers in processing survival computations, methodological note 10. End Results Evaluation Section, National Cancer Institute.

Estève, J., E. Benhamou, M. Croasdale, and L. Raymond. 1990. Relative survival and the estimation of net survival: Elements for further discussion. *Statistics in Medicine* 9: 529–538.

Lambert, P. C., and P. Royston. 2009. Further development of flexible parametric models for survival analysis. *Stata Journal* 9: 265–290.

Nelson, C. P., P. C. Lambert, I. B. Squire, and D. R. Jones. 2007. Flexible parametric models for relative survival, with application in coronary heart disease. *Statistics in Medicine* 26: 5486–5498.

Pohar Perme, M., J. Stare, and J. Estève. 2012. On estimation in relative survival. *Biometrics* 68: 113–120.

About the authors

Michael Hills was a senior lecturer at the London School of Hygiene and Tropical Medicine; he is now retired.

Bernard Rachet is a clinical reader in cancer epidemiology at the London School of Hygiene and Tropical Medicine. His recent research work has focused mainly on the investigation of inequalities in cancer survival and on methods used in population-based cancer survival.

Milena Falcaro is a research fellow in biostatistics and programming at the London School of Hygiene and Tropical Medicine. Her research interests include latent-variable modeling, survival-data analysis, and methods for longitudinal studies.