Estimating Crude Probability of Death in Period Analysis – Technical details of implementation in `strs`

Enzo Coviello
ASL BT
Barletta, Italy
enzo.coviello@tin.it

Ron Dewar
Cancer Care Nova Scotia
Halifax (NS), Canada
Ron.Dewar@ccns.nshealth.ca

Paul W. Dickman
Karolinska Institutet
Stockholm, Sweden
paul.dickman@ki.se

1 Abstract

The Stata command `strs` has for many years supported the estimation of relative survival using both a cohort and period approach, and also estimates crude probabilities of death using a cohort approach. The approach for actuarial estimation of crude probabilities of death in a relative survival framework (Cronin and Feuer, 2000) was developed for cohort analysis and does not directly translate to period analysis. This document describes how we have estimated crude probabilities of death in period analysis. The estimates and confidence intervals are compared to those obtained from a model-based approach and the confidence intervals are compared to those obtained from a bootstrap analysis.

2 Introduction

The primary aim of the user-written Stata command `strs` is to estimate relative survival using a life table approach. For cohort life tables, `strs` employs the usual actuarial estimator; interval-specific observed survival for interval i is $p_i = (1 - d_i/n_i^*)$ where d_i is the number of deaths in the interval i, and $n_i^* = n_i - w_i/2$ the 'effective number at risk' (n_i is the number alive at the start of the interval and w_i the number censored during the interval). In period analysis ([1, 2]), survival times can be left truncated in addition to being right censored, so fewer subjects are at risk for the full interval. In this case, w_i would need to represent the number of individuals whose survival time was left truncated or right censored. Rather than redefining w_i, `strs` estimates survival by transforming the estimated cumulative hazard ($S = \exp(-\Lambda)$) whenever late entry is detected (e.g., if a period approach is employed).

The complement of relative survival (i.e., $1 - RS$) is an estimate of the net probability of death due to cancer – the probability of dying of cancer in a hypothetical world where cancer is the only possible cause of death. This is a distinctly different quantity to the crude probability (CP) of death (also called the cumulative incidence) which represents the probability of dying of cancer (or other causes) in the real world where death due to any cause is possible. Cronin and Feuer [4] showed how crude probabilities of death can be estimated in a relative relative survival framework and their approach is implemented in `strs` for cohort analysis (version 1.2.8, June 2008).

When period analysis is applied, strs uses the so-called hazard transformed (HT) approach. Point estimates of CP can be computed, but the variance-covariance matrix is not immediately available. The variance-covariance matrix described by Cronin and Feuer [4] requires the effective number at risk, n_i^*, which is not available when period analysis is used. The aim of this paper is to show how the variance of the CP of death can be estimated in period analysis. The new approach has been implemented in strs, version 1.4.0.

3 Variance of the observed (all-cause) survival

Before describing how we calculated the variance of the crude probability, we will describe how we calculate the variance of the survivor function for both the actuarial and hazard-transformation approaches since understanding these is central to understanding how we calculated the variance of the crude probability.

Actuarial approach
Applying the method described by Greenwood (1926) [5], the variance of the cumulative observed survival proportion (S_i) up until the end of interval i is given by

$$\mathrm{Var}(S_i) = S_i^2 \left[\sum_{j=1}^{i} \frac{d_j}{n_j^*(n_j^* - d_j)} \right]. \tag{1}$$

For a single interval, Equation 1 reduces to

$$\mathrm{Var}(p_i) = p_i^2 \left\{ \frac{d_i}{n_i^*(n_i^* - d_i)} \right\} = p_i(1 - p_i)/n_i^*,$$

which is the familiar binomial formula for the variance of the interval-specific survival proportion based on n_i* trials. It can also be shown that, in the absence of censoring, Equation 1 reduces to the usual variance of the binomial distribution.

Hazard transformation approach
Applying equation 2.2 in Breslow and Day (1987) [3], the variance of the cumulative hazard Λ_i at the end of interval i is

$$\mathrm{var}(\Lambda_i) = \sum_{j=1}^{i} k_j^2 d_j / y_j^2$$

where k_j is the interval width, d_j is the number of deaths, and y_j is person-time.

By the delta method, the variance of the cumulative survival at the end of interval i is given by

$$\begin{aligned}\mathrm{var}(S_i) &= \mathrm{var}(\exp(-\Lambda_i))\\ &= [\frac{d_i}{d_i\Lambda_i}\exp(-\Lambda_i)]^2\mathrm{var}(\Lambda_i)\\ &= S_i^2\mathrm{var}(\Lambda_i)\\ &= S_i^2\left[\sum_{j=1}^{i}k_j^2 d_j/y_j^2\right]. \qquad (2)\end{aligned}$$

4 Variance of the crude probability of death

Actuarial approach
Expressions for the variance and covariance of the crude probability of death were derived by Cronin and Feuer [4] and are implemented in `strs`.

Hazard transformation approach
The variance of the crude probability of for the actuarial approach contains the term $\frac{d_i}{(n_i^*-d_i)n_i^*}$, which cannot be applied when using the hazard transformation approach since n_i^* is not available. Noting the similarity between Equations 1 and 2, we substituted this term with $(\texttt{end}_i - \texttt{start}_i)^2 d_i/y_i$ where $(\texttt{end}_i - \texttt{start}_i)$ is the length of the interval. In the `strs` code, this term is named `var_Lambda`.

Therefore, in the hazard-transformation approach, we can approximate n_i^* by assuming `var_Lambda` is equivalent to $\frac{d_i}{(n_i^*-d_i)n_i^*}$:

$$\texttt{var\_Lambda}_i = \frac{d_i}{(n_i^*-d_i)n_i^*}$$

$$n_i^*(n_i^*-d_i)\texttt{var\_Lambda}_i - d_i = 0$$

$$n_i^{*2}\texttt{var\_Lambda}_i - n_i^* d_i \texttt{var\_Lambda}_i - d_i = 0$$

$$n_i^* = \texttt{var\_Lambda}_i d_i \pm \frac{\sqrt{(\texttt{var\_Lambda}_i d_i)^2 + 4\texttt{var\_Lambda}_i d_i}}{2\texttt{var\_Lambda}_i)}$$

Therefore, when estimating crude probabilities of death using a hazard-transformation approach (i.e., for period analysis) we use the expressions for the variance and covariance described by Cronin and Feuer [4] but with n_i^* calculated as above (using + operator).

5 Examples

The first example shows how crude probabilities are estimated in cohort analysis by applying the Cronin and Feuer approach already implemented in `strs`. We use month intervals but for sake of brevity show only estimates at the end of each follow-up year.

```
. use colon_net, clear
(Colon carcinoma, Finland 1975-94, follow-up to 1995. Dates continuous.)
. quietly stset exit, origin(dx) f(status) scale(365.24) id(id)
. strs using popmort, br(0(.08333333)10) mergeby(_year sex _age) ///
>         cuminc format(%10.4f) list(cr_e2 ci_dc lo_ci_dc hi_ci_dc ci_do lo_ci_do hi_ci_do)
         failure _d:  status
   analysis time _t:  (exit-origin)/365.24
             origin:  time dx
                 id:  id
No late entry detected - p is estimated using the actuarial method

  start end   cr_e2   ci_dc   lo_ci_dc   hi_ci_dc    ci_do  lo_ci_do  hi_ci_do

  .9167   1  0.6607  0.3345     0.3267     0.3424   0.0351    0.0348    0.0353
  1.917   2  0.5675  0.4216     0.4132     0.4300   0.0620    0.0614    0.0626
  2.917   3  0.5231  0.4611     0.4523     0.4699   0.0859    0.0850    0.0869
  3.917   4  0.4946  0.4852     0.4760     0.4943   0.1083    0.1070    0.1096
  4.917   5  0.4734  0.5022     0.4927     0.5116   0.1293    0.1276    0.1310

  5.917   6  0.4584  0.5135     0.5037     0.5233   0.1490    0.1469    0.1511
  6.917   7  0.4487  0.5205     0.5104     0.5306   0.1677    0.1652    0.1701
  7.917   8  0.4456  0.5227     0.5121     0.5331   0.1856    0.1828    0.1885
  8.917   9  0.4399  0.5263     0.5154     0.5371   0.2029    0.1997    0.2062
  9.917  10  0.4365  0.5283     0.5170     0.5395   0.2198    0.2161    0.2234
```

In the output above, 1 minus `cr_e2` is the net probability of death due to cancer while `ci_dc` and `ci_do` are the crude probabilities of death (also known as cumulative incidence) due to cancer and other causes respectively. Lower (`lo_ci`) and upper (`hi_ci`) limits for 95% confidence intervals are also provided.

In cohort analysis, we can force `strs` to use the HT approach by specifying the option `ht`. The next table shows that the CP of death (and corresponding 95% confidence intervals) estimated by using the HT approach are strictly comparable with the previous ones where the Cronin and Feuer method is applied.

```
. strs using popmort, br(0(.08333333)10) mergeby(_year sex _age) ///
>         cuminc format(%10.4f) ht ///
>         list(cr_e2 ci_dc lo_ci_dc hi_ci_dc ci_do lo_ci_do hi_ci_do)
         failure _d:  status
   analysis time _t:  (exit-origin)/365.24
             origin:  time dx
                 id:  id
The conditional survival proportion (p) is estimated by transforming the
estimated cumulative hazard rather than by the actuarial method (default for cohort analysis).
See http://pauldickman.com/rsmodel/stata_colon/standard_errors.pdf for details.

  +-----------------------------------------------------------------------------+
  | start end    cr_e2    ci_dc   lo_ci_dc  hi_ci_dc    ci_do  lo_ci_do  hi_ci_do |
  |-----------------------------------------------------------------------------|
  | .9167   1   0.6590   0.3362     0.3284    0.3440   0.0350    0.0347    0.0353 |
  | 1.917   2   0.5661   0.4231     0.4146    0.4315   0.0619    0.0613    0.0625 |
  | 2.917   3   0.5218   0.4625     0.4536    0.4713   0.0857    0.0848    0.0867 |
  | 3.917   4   0.4933   0.4865     0.4773    0.4956   0.1081    0.1068    0.1094 |
  | 4.917   5   0.4721   0.5035     0.4940    0.5129   0.1290    0.1273    0.1307 |
  |-----------------------------------------------------------------------------|
  | 5.917   6   0.4572   0.5148     0.5050    0.5245   0.1486    0.1466    0.1507 |
  | 6.917   7   0.4475   0.5218     0.5116    0.5318   0.1672    0.1648    0.1697 |
  | 7.917   8   0.4444   0.5239     0.5134    0.5343   0.1851    0.1823    0.1880 |
  | 8.917   9   0.4387   0.5275     0.5166    0.5383   0.2024    0.1992    0.2057 |
  | 9.917  10   0.4354   0.5296     0.5183    0.5407   0.2192    0.2155    0.2229 |
  +-----------------------------------------------------------------------------+
```

It is also possible to estimate crude probabilities of death from model-based estimates of relative survival, rather than life table estimates of relative survival[6]. In Stata, this can be done (for either cohort or period estimation) after fitting a flexible parametric model using `stpm2` and the post-estimation command `stpm2cm` [7]. The model-based estimates are for specific values of covariates, such as age, sex, year of diagnosis, whereas the life table approach provides marginal estimates for all patients included in the life table. To produce life table CP estimates comparable with those produced by a flexible parametric model, we assign all patients the same sex and age at diagnosis. Furthermore, in the period analysis the time at risk spans the 1985 calendar year. Commands used for obtaining life table estimates in such a hypothetical scenario are listed below. Figure 1 shows that the agreement between the life table and model-based estimates is fairly good.

```
. use colon_net, clear
(Colon carcinoma, Finland 1975-94, follow-up to 1995. Dates continuous.)

. replace age = 65
(15563 real changes made)

. replace sex=1
(9224 real changes made)

. stset exit, origin(dx) enter(time mdy(1,1,1985)) f(status) ///
>         id(id) scale(365.24) exit(time mdy(12,31,1985))

                id:  id
     failure event:  status != 0 & status < .
obs. time interval:  (exit[_n-1], exit]
 enter on or after:  time mdy(1,1,1985)
 exit on or before:  time mdy(12,31,1985)
    t for analysis:  (time-origin)/365.24
            origin:  time dx
------------------------------------------------------------------------------
    15564  total observations
     3820  observations end on or before enter()
     8284  observations begin on or after exit
------------------------------------------------------------------------------
     3460  observations remaining, representing
     3460  subjects
      564  failures in single-failure-per-subject data
  2783.72  total analysis time at risk and under observation
                                                at risk from t =         0
                                     earliest observed entry t =         0
                                          last observed exit t =  10.99003

. strs using popmort, br(0(.08333333)10) mergeby(_year sex _age) cuminc ///
>         savgr(checkcuminclatestpm,replace)
  (output omitted)
```

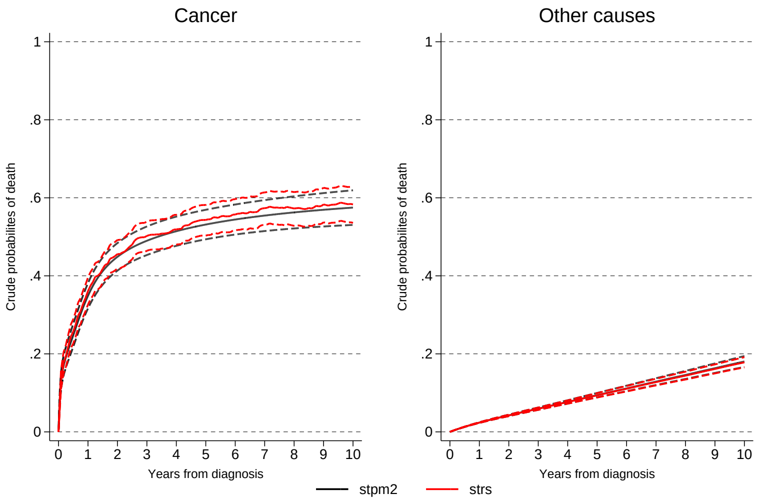

Figure 1: Flexible Parametric and life table CP estimates (with 95% CI) for cancer and other causes of death

Finally we compare the confidence intervals obtained using the method implemented in `strs` with those using bootstrapping with 1000 replications. The bias corrected method was used to calculate the boostrapped confidence intervals.

First we compare the two methods in the cohort analysis where `strs` applies the original formula of the variance-covariance matrix proposed by Cronin and Feuer. Cases diagnosed from 1980 up to 1984 were selected from the colon_net dataset. Graphs in the left side of figure 2 show that for the CP of death due to cancer `strs` slightly overestimates the upper bound and underestimates the lower bound of the confidence intervals computed using bootstrapping. The opposite happens when the CP of death due to other causes is considered. Then, we compare the two methods in the period analysis where `strs` applies the proposed solution. This analysis considers the survival experience of cases observed in the time window starting from the beginning of 1990 and ending at the end of 1994. Results of this comparison are shown in the graphs on the right side of the figure 2. They are very similar to the previous ones showing that in the period analysis `strs` and bootstrap confidence intervals agree as well as in the cohort analysis.

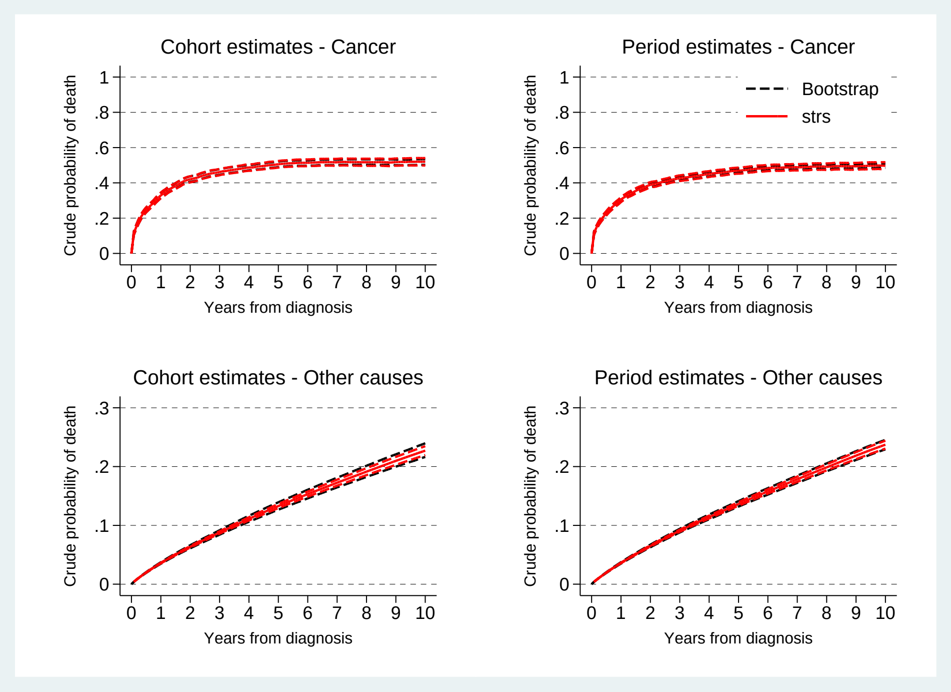

Figure 2: Comparison of 95% confidence intervals for the CP of death due to cancer and other causes using the Cronin and Feuer formula (cohort analysis, left side, black dashed line), the proposed approach (period analysis, right side, black dashed line) and bootstrapping (red dashed line).

6 Conclusion

Period analysis exploits the most recent survival experience of the cancer cases to produce up-to-date estimates of their survival probability. The approach we proposed, implemented in the version 1.4.0 of `strs`, extends the application of the period analysis to the estimation of the crude probabilities of death due to cancer and other causes.

References

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

[2] H. Brenner and B. Rachet. Hybrid analysis for up-to-date long-term survival rates in cancer registries with delayed recording of incident cases. *European Journal of Cancer*, 40(16):2494–2501, 2004.

[3] N. E. Breslow and N. E. Day. *Statistical Methods in Cancer Research: Volume II - The Design and Analysis of Cohort Studies.* IARC Scientific Publications No. 82. Lyon: IARC, 1987.

[4] K. A. Cronin and E. J. Feuer. Cumulative cause-specific mortality for cancer patients in the presence of other causes: a crude analogue of relative survival. *Statistics in Medicine*, 19(13):1729–1740, 2000.

[5] M. Greenwood. *The Errors of Sampling of the Survivorship Table*, volume 33 of *Reports on Public Health and Medical Subjects.* London: Her Majesty's Stationery Office, 1926.

[6] P. C. Lambert, P. W. Dickman, C. P. Nelson, and P. Royston. Estimating the crude probability of death due to cancer and other causes using relative survival models. *Stat Med*, 29:885 – 895, 2010.

[7] P. C. Lambert and P. Royston. Further development of flexible parametric models for survival analysis. *Statistics in Medicine*, 9:265–290, 2009.