Flexible parametric survival models on the log hazard scale: The `strcs` command

Hannah Bower*
Michael J. Crowther and Paul C. Lambert

*Department of Medical Epidemiology and Biostatistics
Karolinska Institutet, Sweden

Nordic and Baltic Stata Users Group meeting,
4th September 2015

Outline

1. Introduction
2. Flexible parametric survival models
3. The `strcs` command
4. Examples
5. Conclusions
6. References

Introduction

- Cox is the most widely used survival model [Cox, 1972]
- Parametric models are also implemented frequently, flexible parametric survival models are becoming more popular [Royston and Lambert, 2011]
- `stgenreg` fits parametric models with user-defined hazards [Crowther and Lambert, 2013]
- `strcs` is an extension to `stgenreg` when one wants to model the hazard function using restricted cubic splines

Flexible parametric survival models

- Flexible parametric survival models (FPSMs) use restricted cubic splines (RCS) to model some form of the hazard function
- RCS are piecewise cubic polynomials joined together at points called knots
  - Continuous 1st, and 2nd derivatives at the knots, linear before first and after last knot
- RCS are able to capture complex hazard functions which standard parametric models may struggle to capture

Flexible parametric survival models

- We usually fit FPSMs on the log cumulative hazard scale
- FPSM on the log cumulative hazard scale can be written as:

$$\ln(H(t;\boldsymbol{x})) = \underbrace{s(\ln(t);\gamma_{\mathbf{0}})} + \overbrace{\boldsymbol{x}\beta}$$

Flexible parametric survival models

- We usually fit FPSMs on the log cumulative hazard scale
- FPSM on the log cumulative hazard scale can be written as:

$$\ln(H(t;\boldsymbol{x})) = \underbrace{s(\ln(t);\gamma_{\mathbf{0}})}_{\text{spline function}} + \overbrace{\boldsymbol{x}\beta}$$

Flexible parametric survival models

- We usually fit FPSMs on the log cumulative hazard scale
- FPSM on the log cumulative hazard scale can be written as:

$$\ln(H(t;\boldsymbol{x})) = \underbrace{s(\ln(t);\gamma_{\mathbf{0}})}_{\text{spline function}} + \overbrace{\boldsymbol{x}\beta}^{\text{covariates}}$$

Flexible parametric survival models

- We usually fit FPSMs on the log cumulative hazard scale
- FPSM on the log cumulative hazard scale can be written as:

$$\ln(H(t;\boldsymbol{x})) = \underbrace{s(\ln(t);\gamma_{\mathbf{0}})}_{\text{spline function}} + \overbrace{\boldsymbol{x}\beta}^{\text{covariates}} + \underbrace{\sum_{k=1}^{D} s(\ln(t);\gamma_{\boldsymbol{k}})\boldsymbol{x_k}}_{\text{time-dependent effects}}$$

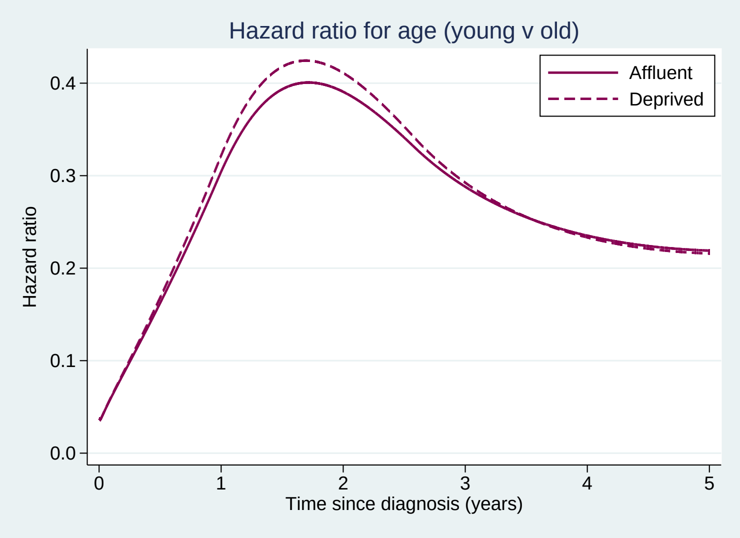
Flexible parametric survival models

- `stpm2` fits FPSMs on the log cumulative hazard scale in Stata [Lambert and Royston, 2009]
- Cumulative hazard shape is easier to capture
- It is computationally intensive to fit models on the log hazard scale
- However, we have problems when we have multiple time-dependent effects on the log cumulative hazard scale

The problem with multiple time-dependent effects

- 14,423 women diagnosed with breast cancer in England and Wales [Coleman et al., 1999]
  - `young`: $<$50 years or 80+ years at diagnosis
  - `affluent`: least deprived or most deprived
- Fit a FPSM on the log cumulative hazard scale with time-dependent effects for deprivation and age at diagnosis
  - No interaction between deprivation and age
- Predict the hazard ratio for age in each of the deprivation levels

The problem with multiple time-dependent effects



The log hazard scale

- Non-proportional FPSM on the log hazard scale:

$$\ln(h(t;\boldsymbol{x})) = \underbrace{s(\ln(t);\gamma_{\boldsymbol{0}})}_{\text{spline function}} + \overbrace{\boldsymbol{x}\beta}^{\text{covariates}} + \underbrace{\sum_{k=1}^{D} s(\ln(t);\gamma_{\boldsymbol{k}})\boldsymbol{x}_{\boldsymbol{k}}}_{\text{time-dependent effects}}$$

The log hazard scale

- Non-proportional FPSM on the log hazard scale:

$$\ln(\mathbf{h}(t;\boldsymbol{x})) = \underbrace{s(\ln(t);\gamma_{\mathbf{0}})}_{\text{spline function}} + \overbrace{\boldsymbol{x}\beta}^{\text{covariates}} + \underbrace{\sum_{k=1}^{D} s(\ln(t);\gamma_{\boldsymbol{k}})\boldsymbol{x}_{\boldsymbol{k}}}_{\text{time-dependent effects}}$$

Maximum likelihood estimation

Log-likelihood

$$\log L_i = d_i \log\{h(t_i)\} - H(t_i)$$

- d_i = event indicator
- $h(t_i)$ = hazard function
- $H(t_i)$ = cumulative hazard function

$$H(t_i) = \int_0^t h(u_i)du$$

Maximum likelihood estimation

Log-likelihood

$$\log L_i = d_i \log\{h(t_i)\} - H(t_i)$$

- **FPSMs on the log cumulative hazard**: analytically differentiate to get hazard function
- **FPSMs on the log hazard scale**: numerical integration required to get cumulative hazard function

Gaussian quadrature

- Gaussian quadrature converts an integral of some hazard function $h(x)$ into a weighted summation over a set of pre-defined points known as nodes

$$\int_{t_0}^{t} h(z)dz \approx \frac{t - t_0}{2} \sum_{j=1}^{m} w_j h(\frac{t - t_0}{2} z_j + \frac{t_0 + t}{2}) \tag{1}$$

where m and z_j represent the number of nodes and the node locations, respectively.

The `strcs` command

- `strcs` is a Stata command which fits FPSMs on the log hazard scale
- Integration of the hazard is performed in two steps [Crowther and Lambert, 2014]:
  1. Analytical integration before the first, and after the last knot
  2. Gauss-Legendre quadrature numerical integration in between the first and last knot
- This reduces the number of nodes required and thus the computational intensity
- `stgenreg` performs numerical integration over the whole function since it is a general tool

The `strcs` command

- `strcs` is a Stata command which fits FPSMs on the log hazard scale
- Integration of the hazard is performed in two steps [Crowther and Lambert, 2014]:
  1. Analytical integration before the first, and after the last knot
  2. Gauss-Legendre quadrature numerical integration in between the first and last knot
- This reduces the number of nodes required and thus the computational intensity
- `stgenreg` performs numerical integration over the whole function since it is a general tool

`strcs` syntax

`strcs` [*varlist*], `df(#)` [`tvc`(*varlist*) ...]

- `df(#)` - defines degrees of freedom for baseline
- `tvc`(*varlist*) - defines covariates with time-dependent effects
- `dftvc`(*df_list*) - defines the degrees of freedom of time-dependent effects
- `nodes(#)` - defines the number of nodes used within numerical integration
- `bhazard`(*varname*) - invokes relative survival models
- Other options: smooth baseline hazard over time, specify knot positions, ...

Example: Proportional hazards model

```
. strcs affluent young, df(3)

Log likelihood = -17610.978                     Number of obs   =      14423

------------------------------------------------------------------------------
             | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
    affluent |   .8412791   .0216063    -6.73   0.000      .7999797    .8847108
       young |   .2357943   .0060132   -56.65   0.000      .2242983    .2478795
-------------+----------------------------------------------------------------
rcs          |
        __s1 |  -.2417658   .0140943   -17.15   0.000       -.26939   -.2141415
        __s2 |  -.0837641   .0122397    -6.84   0.000     -.1077536   -.0597747
        __s3 |   .0106206   .0113675     0.93   0.350     -.0116593    .0329006
       _cons |  -1.149726   .0300179   -38.30   0.000      -1.20856   -1.090892
------------------------------------------------------------------------------
```

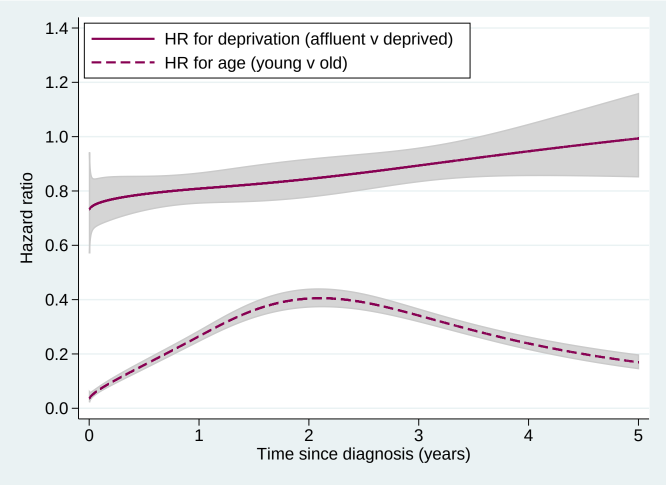
Example: Non-proportional hazards model

```
. strcs affluent young, df(3) tvc(affluent young) dftvc(3)

Log likelihood =  -17387.46                        Number of obs    =      14423
------------------------------------------------------------------------------
             | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
    affluent |   .9231825   .0447397    -1.65   0.099     .8395299     1.01517
       young |   .1899977   .0089422   -35.29   0.000     .1732554    .2083579
-------------+----------------------------------------------------------------
rcs          |
        __s1 |  -.3222852   .0257495   -12.52   0.000    -.3727533   -.2718171
        __s2 |  -.1464735   .0227574    -6.44   0.000    -.1910771   -.1018699
        __s3 |  -.1021786   .0206593    -4.95   0.000      -.14267   -.0616872
__s_affluent1 |   .0862308   .0294077     2.93   0.003     .0285927    .1438688
__s_affluent2 |  -.0418096   .0257252    -1.63   0.104    -.0922301    .0086109
__s_affluent3 |  -.0196183   .0237653    -0.83   0.409    -.0661974    .0269608
  __s_young1 |   .2162942    .034253     6.31   0.000     .1491596    .2834288
  __s_young2 |   .2530733    .028519     8.87   0.000     .1971771    .3089694
  __s_young3 |   .2960521   .0248382    11.92   0.000     .2473702    .3447341
       _cons |    -1.1463   .0438901   -26.12   0.000    -1.232323   -1.060277
------------------------------------------------------------------------------
```

Example: Non-proportional hazards model

```
. predict hr_affluent, hrnumerator(affluent 1) hrdenominator(affluent 0) ci
. predict hr_young, hrnumerator(young 1) hrdenominator(young 0) ci
```



Other post-estimation predictions

- Survival function
- Differences in survival functions between groups
- Hazard function
- Differences in hazard functions between groups
- Cumulative hazard function

Conclusions

- Fitting FPSMs on the log hazard scale using `strcs` is an alternative to fitting FPSMs on the log cumulative hazard scale
- Use `strcs` if you have many time-dependent effects and wish to present HRs for covariates
- The need for numerical integration slows things down
- Nodes may need to be increased, may need sensitivity analyses
- Require fewer nodes than `stgenreg` due to two-step integration process

References I

[Coleman et al., 1999] Coleman, M. P., Babb, P., Damiecki, P., Grosclaude, P., Honjo, S., Jones, J., Knerer, G., Pitard, A., Quinn, M., Sloggett, A., and De Stavola, B. (1999).
Cancer Survival Trends in England and Wales, 1971–1995: Deprivation and NHS Region.
Number 61 in Studies in Medical and Population Subjects. London: The Stationery Office.

[Cox, 1972] Cox, D. R. (1972).
Regression models and life-tables (with discussion).
JRSSB, 34:187–220.

[Crowther and Lambert, 2013] Crowther, M. J. and Lambert, P. C. (2013).
stgenreg: A stata package for general parametric survival analysis.
Journal of Statistical Software, 53:1–17.

[Crowther and Lambert, 2014] Crowther, M. J. and Lambert, P. C. (2014).
A general framework for parametric survival analysis.
Stat Med, 33(30):5280–5297.

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

[Royston and Lambert, 2011] Royston, P. and Lambert, P. C. (2011).
Flexible parametric survival analysis in Stata: Beyond the Cox model.
Stata Press.