

# Using Spline Functions on the Log Cumulative Hazard scale, the STRSRCS Command

Christopher P Nelson  
Center for Biostatistics and Genetic Epidemiology  
University of Leicester  
cn46@le.ac.uk

Methods for studying cancer patient survival, with application in Stata:  
September 6 2007



## Contents

- 1 Introduction to the Methods
- 2 Example code
- 3 Survival estimates
- 4 Excess hazard estimates
- 5 Excess hazard rate ratios
- 6 Comparison with piecewise model
- 7 Modelling continuous covariates
- 8 Issues

## Issues with current methods

- Current methods in practice split the timescale
- Piecewise models could be considered as biologically implausible
- Many parameters for time dependent effects
- Collapsed data does not allow modelling of continuous covariates
- Can lead to problems in *small datasets*, e.g. *zero cells*
- Current spline methods that do not split time use numerical integration
  - this slows the process
  - Giorgi et al. (2003)
  - Bolard et al. (2002)

## A difference in scale

- Current relative survival models are fitted to the **log excess hazard** scale, i.e.

$$h(t) = h^*(t) + \exp(\beta z)$$

- $\beta$  = a vector of parameters to be estimated for covariates  $z$
- whereas we are interested in modelling on the **log cumulative excess hazard** scale, i.e.

$$H(t) = H^*(t) + \Lambda(t)$$

$$\ln(-\ln R(t; z)) = \ln(\Lambda(t)) = \ln(\Lambda_0(t)) + \beta z$$

- $H(t)$  = cumulative overall hazard,  $H^*(t)$  = cumulative expected hazard,  $\Lambda(t)$  = cumulative excess hazard &  $\Lambda_0(t)$  = baseline cumulative excess hazard
- Note that proportional cumulative excess hazards implies proportional excess hazards

# A Flexible Parametric Model for Survival analyses

## Royston and Parmar (2002)

- Time by covariate interactions easy (non-PH)
- Ability to model the baseline hazard
- Proportional Hazards ( $\theta \rightarrow 0$ ) or Proportional Odds ( $\theta = 1$ ) models are available
- Further extensions to allow non-proportional effects of some or all of the covariates (other values of  $\theta$ )
- And can be extended to relative survival to give the following likelihood

$$L_i = d_i \ln \left[ h^*(t_i) + \frac{1}{t_i} \frac{ds(x_i; \gamma)}{dx_i} \exp(\eta_i) \right] - \exp(\eta_i)$$

- Hazard and survival are both obtained analytically.

## Stata Code

### Merge expected mortality

```
gen _age = int(min(agehosp + _t,99))
gen _year = int(min(yearhosp + _t,2006))
sort _year sex _age
merge _year sex _age using popmort, nokeep
gen rate = -ln(prob)
```

- Building a model

### PEH model

```
xi: strsrcs i.year8594 i.sex i.agegrp, bhaz(rate) scale(hazard) df(4)
```

### Age non-PEH model

```
xi: strsrcs i.year8594 i.sex , bhaz(rate) scale(hazard) df(4) strata(i.agegrp)
```

## PEH output

```

Variables _rcs1 to _rcs4 and _d_rcs1 to _d_rcs4 were created
                                Number of obs =      15564
                                Wald chi2(5)   =      256.69
Log likelihood = -20984.928       Prob > chi2   =      0.0000
-----+-----

```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----						
xb						
_Iyear8594_1	-.1744355	.0248468	-7.02	0.000	-.2231342	-.1257367
_Isex_2	-.00843	.0255469	-0.33	0.741	-.0585009	.041641
_Iagegrp_1	.0730139	.0626038	1.17	0.243	-.0496873	.1957152
_Iagegrp_2	.2006328	.0578447	3.47	0.001	.0872592	.3140064
_Iagegrp_3	.5164176	.0586398	8.81	0.000	.4014858	.6313495
_cons	.1206518	.0843953	1.43	0.153	-.04476	.2860636
-----+-----						
s1						
_cons	1.354656	.0317743	42.63	0.000	1.292379	1.416932
-----+-----						
s2						
_cons	.1464346	.0107499	13.62	0.000	.1253652	.1675039
-----+-----						
s3						
_cons	-.1662767	.0158295	-10.50	0.000	-.197302	-.1352515
-----+-----						
s4						
_cons	.1074269	.0106424	10.09	0.000	.0865682	.1282855
-----+-----						

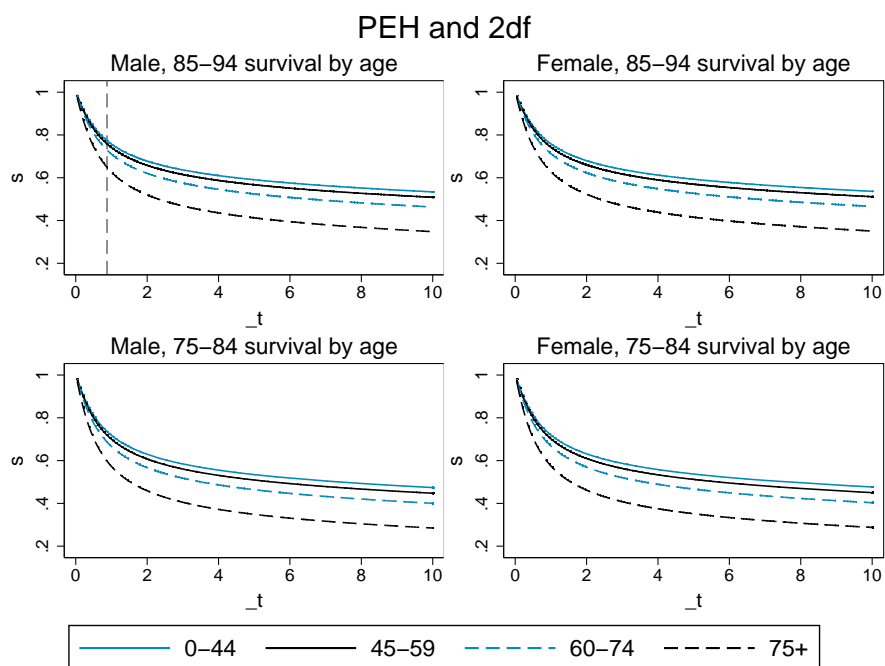
## Example relative survival plot

### Obtaining relative survival estimates

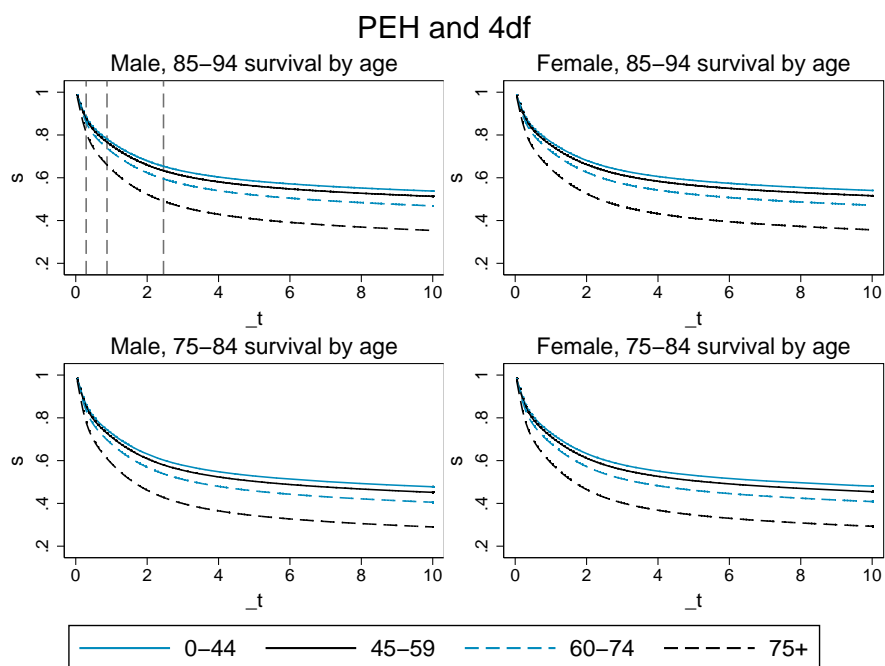
```
predict survival , survival
```

```
note: New variable survival has been created
```

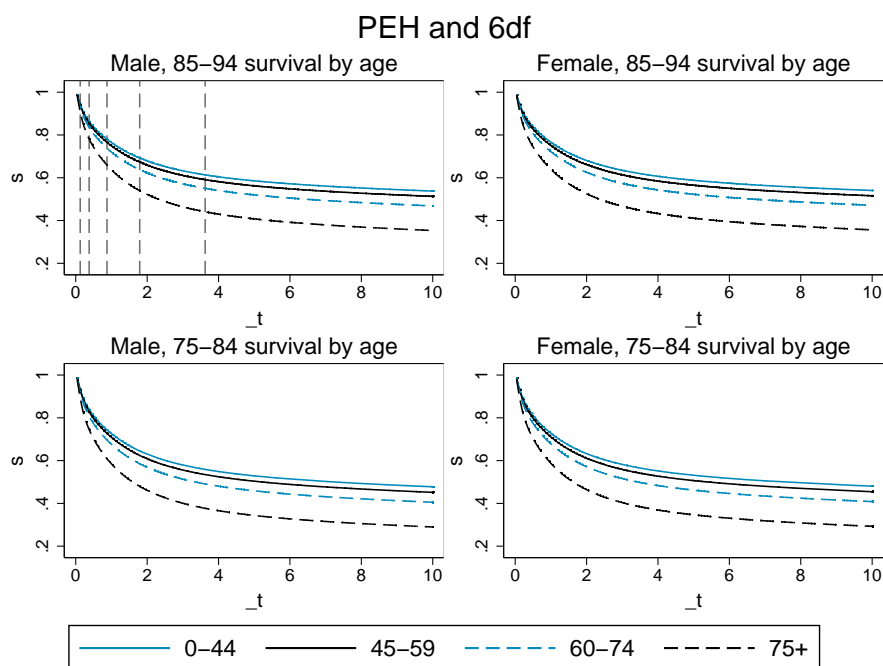
## Example relative survival plot



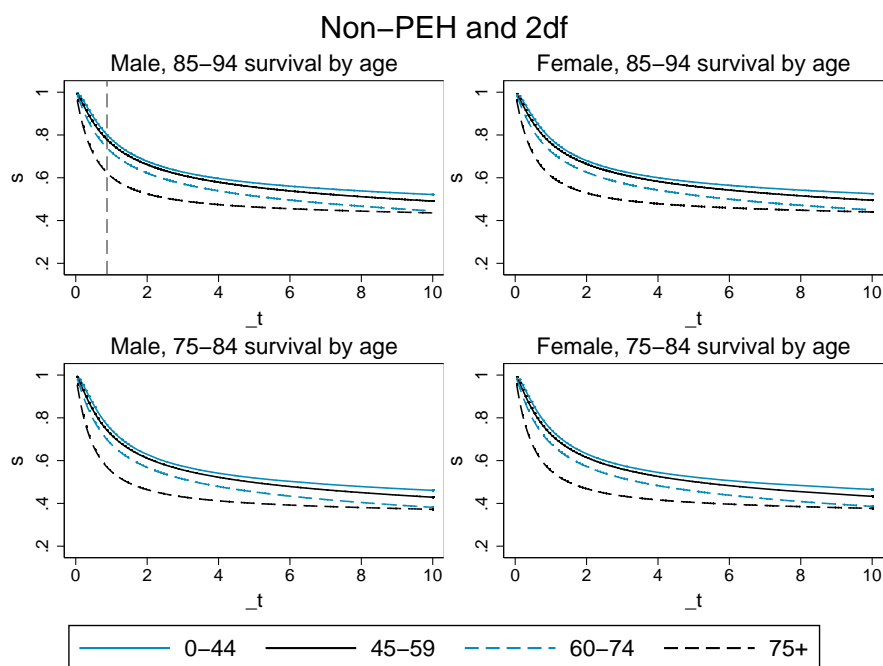
## Example relative survival plot



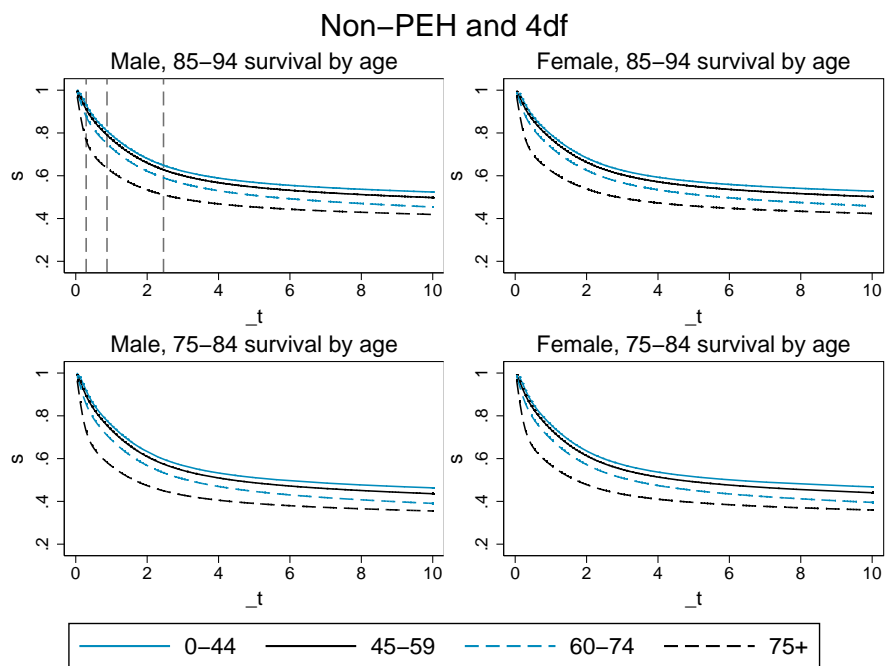
## Example relative survival plot



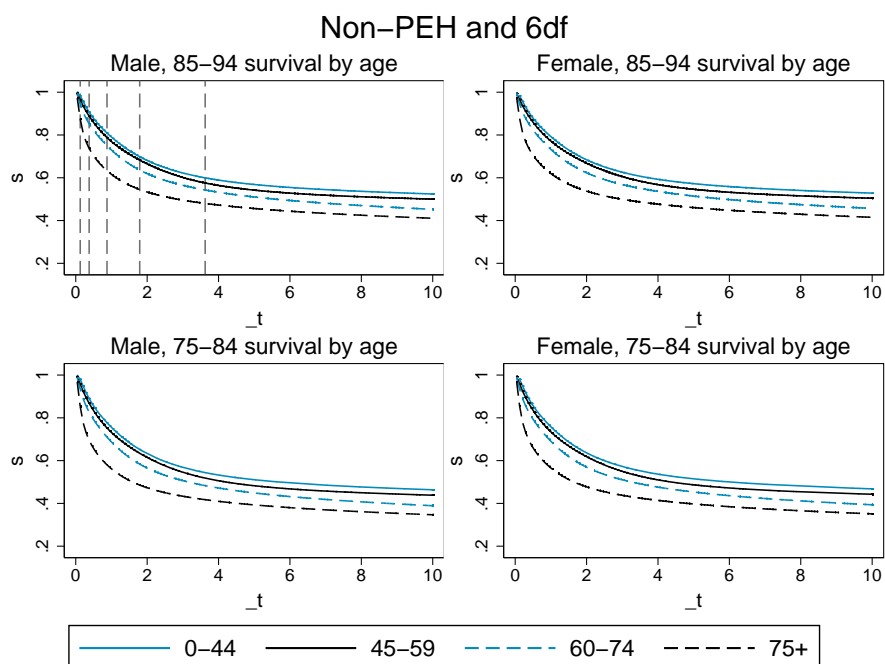
## Example relative survival plot



## Example relative survival plot



## Example relative survival plot



## Example excess mortality plot

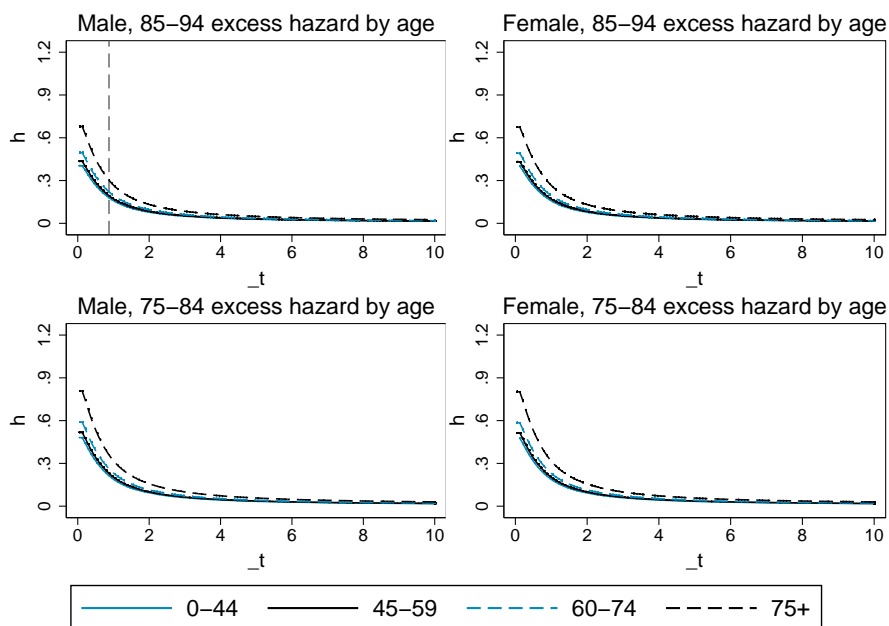
### Obtaining excess mortality estimates

```
predict hazard , hazard
```

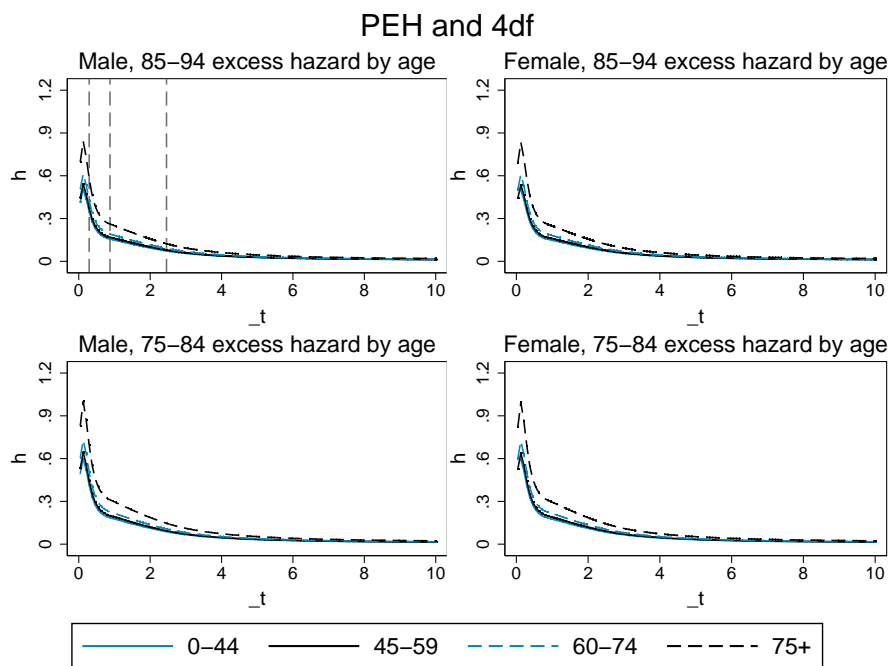
note: New variable hazard has been created

## Example excess mortality plot

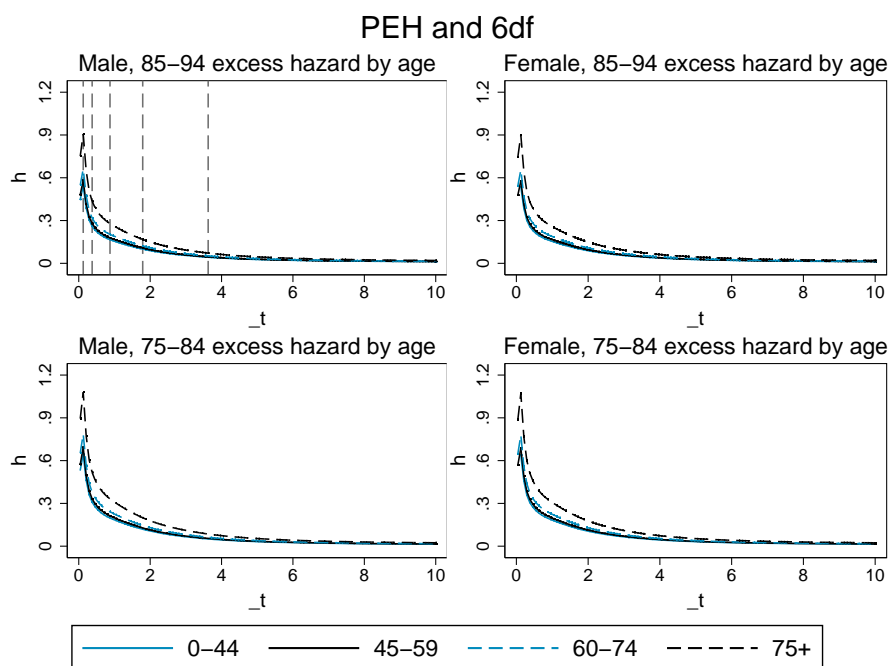
### PEH and 2df



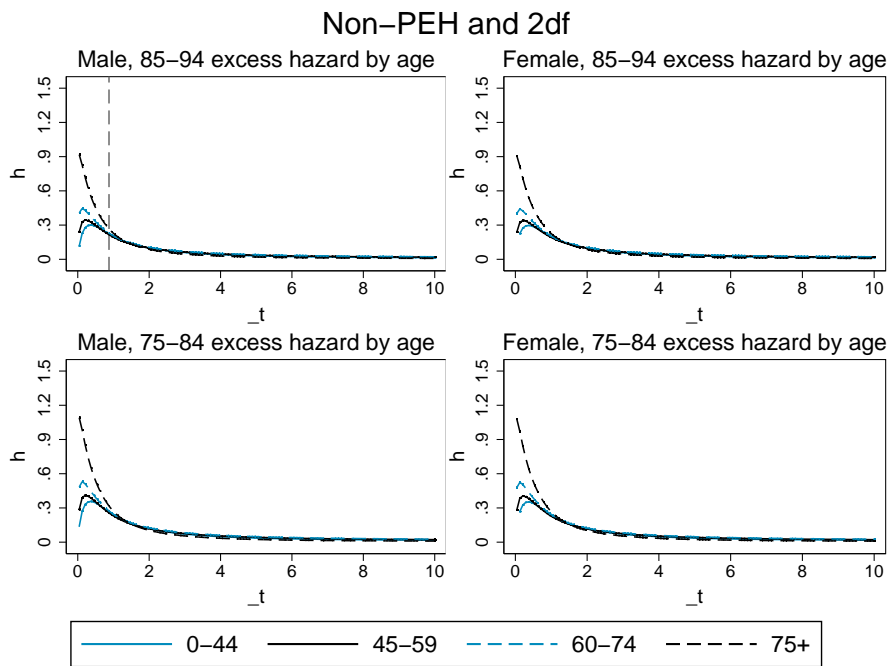
# Example excess mortality plot



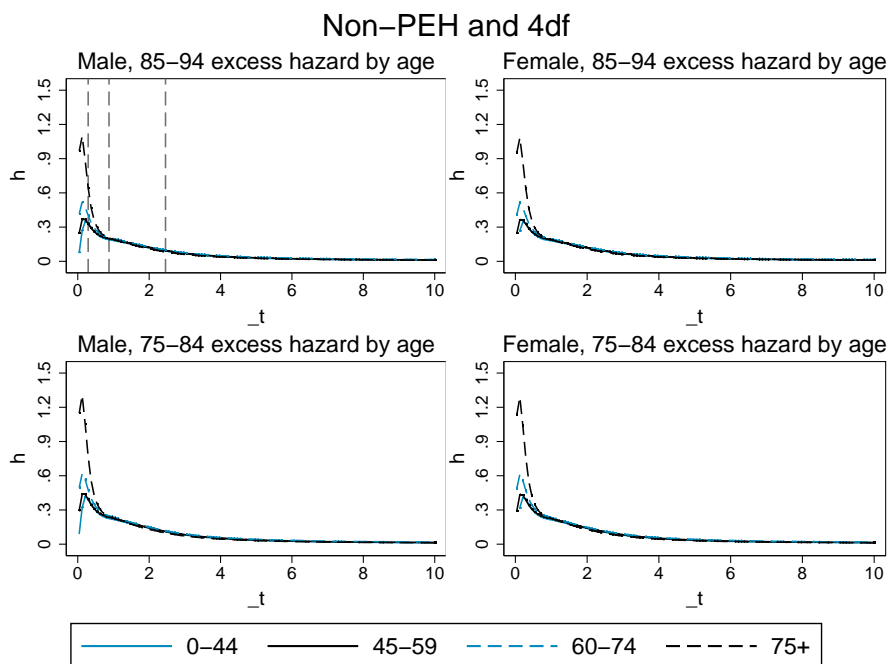
# Example excess mortality plot



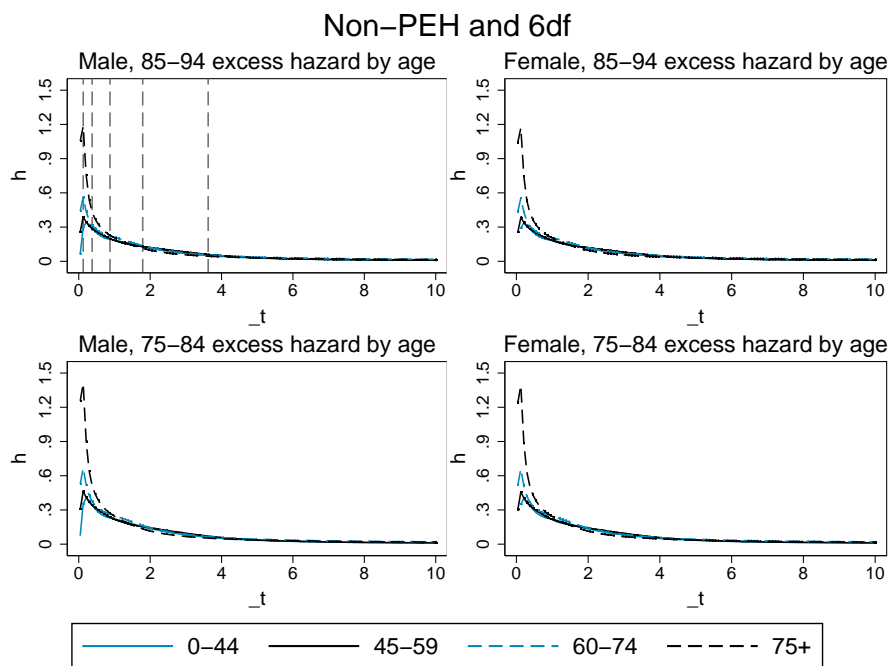
# Example excess mortality plot



# Example excess mortality plot



## Example excess mortality plot



## Model selection

### Using the AIC

AIC		
DF	PEH	non-PEH
2	42172.07*	41936.92
4	41989.86	41720.98
6	41937.64	41670.46†

- \* took 18.0 seconds to run
- † took 76.4 seconds to run

## Excess Hazard Rate Ratios

- Calculating ratios is not easy as it is not currently implemented in the program

### Stata code

```
. program hratio
1.     tempvar lhr1 lhr1_lci lhr1_uci t lhr2 lhr2_lci lhr2_uci lhr3 lhr3_lci lhr3_uci
2.
.     forvalues i=1/'e(df)' ///
>     {
3.         local rcs0 "'rcs0' [s'i'][_cons]*_rcs'i'"
4.         local rcs1 "'rcs1' [s'i'][_cons]*_rcs'i' + [s'i'][_agegp2]*_rcs'i'"
5.         local rcs2 "'rcs2' [s'i'][_cons]*_rcs'i' + [s'i'][_agegp3]*_rcs'i'"
6.         local rcs3 "'rcs3' [s'i'][_cons]*_rcs'i' + [s'i'][_agegp4]*_rcs'i'"
7.
.
.
.     if 'i' != 'e(df)' ///
>     {
8.         local rcs0 "'rcs0' + "
9.         local rcs1 "'rcs1' + "
10.        local rcs2 "'rcs2' + "
11.        local rcs3 "'rcs3' + "
12.    }
...
REPEAT FOR drcs
23. }
```

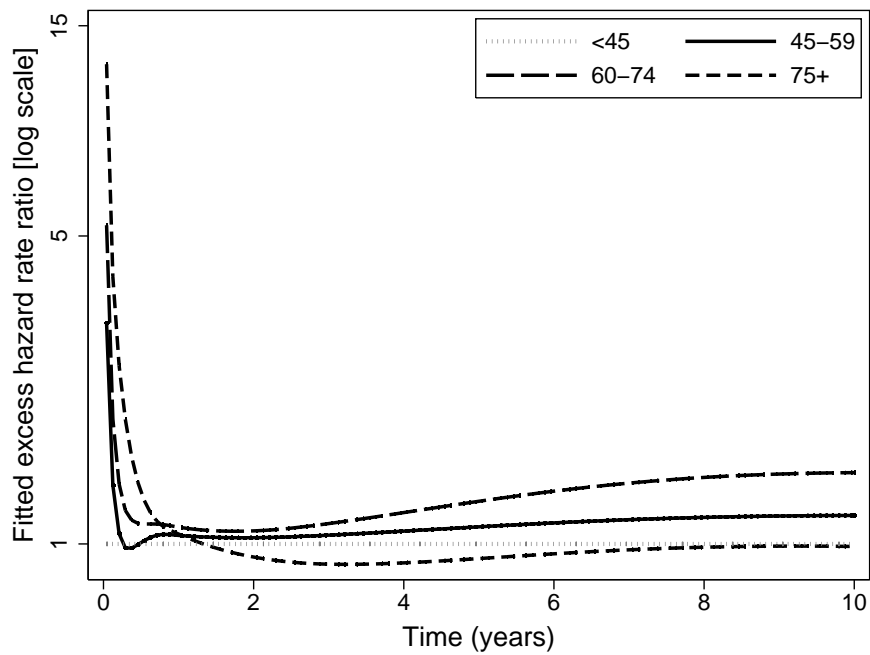
## Excess Hazard Rate Ratios

### Stata code continued

```
.     gen 't'=_t
25.    qui predictnl double 'lhr1' = (-ln('t') + ln('dracs1') + ([xb][_cons] + [xb][_agegp2] + 'rcs1')) - ///
>        (-ln('t') + ln('dracs0') + ([xb][_cons] + 'rcs0')), ci('lhr1_lci' 'lhr1_uci')
26.    qui predictnl double 'lhr2' = (-ln('t') + ln('dracs2') + ([xb][_cons] + [xb][_agegp3] + 'rcs2')) - ///
>        (-ln('t') + ln('dracs0') + ([xb][_cons] + 'rcs0')), ci('lhr2_lci' 'lhr2_uci')
27.    qui predictnl double 'lhr3' = (-ln('t') + ln('dracs3') + ([xb][_cons] + [xb][_agegp4] + 'rcs3')) - ///
>        (-ln('t') + ln('dracs0') + ([xb][_cons] + 'rcs0')), ci('lhr3_lci' 'lhr3_uci')
28.    qui gen double hratio'e(df)'1=exp('lhr1')
29.    qui gen double hratio'e(df)'1_lci=exp('lhr1_lci')
30.    qui gen double hratio'e(df)'1_uci=exp('lhr1_uci')
31.    qui gen double hratio'e(df)'2=exp('lhr2')
32.    qui gen double hratio'e(df)'2_lci=exp('lhr2_lci')
33.    qui gen double hratio'e(df)'2_uci=exp('lhr2_uci')
34.    qui gen double hratio'e(df)'3=exp('lhr3')
35.    qui gen double hratio'e(df)'3_lci=exp('lhr3_lci')
36.    qui gen double hratio'e(df)'3_uci=exp('lhr3_uci')
37.
.     di "note: new variable hratio'e(df)'1 , hratio'e(df)'2 & hratio'e(df)'3 have been created"
38. end
```

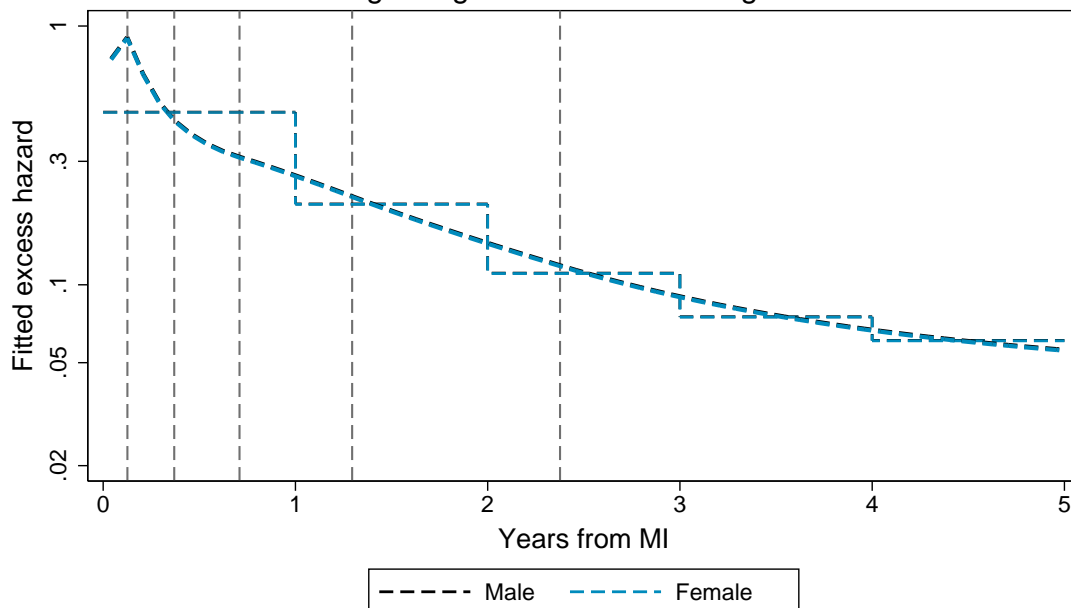
- The new variables can be plotted
- Potential issue when there are two or more time dependent effects

# Hazard Rate Ratios

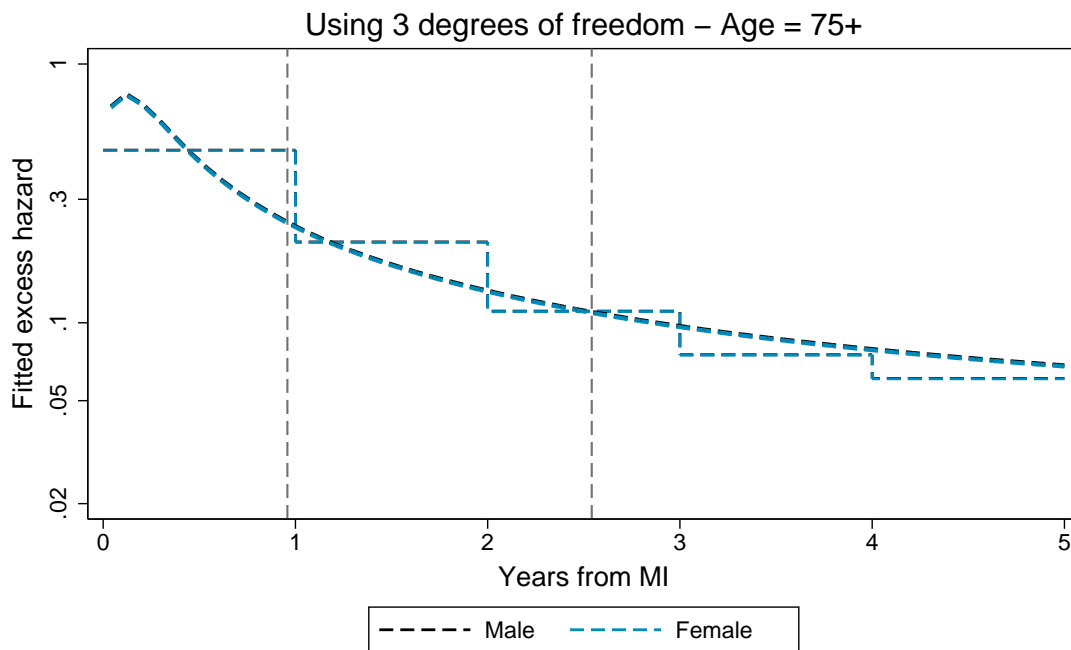


# Compare to GLM, Dickman et al. (2004)

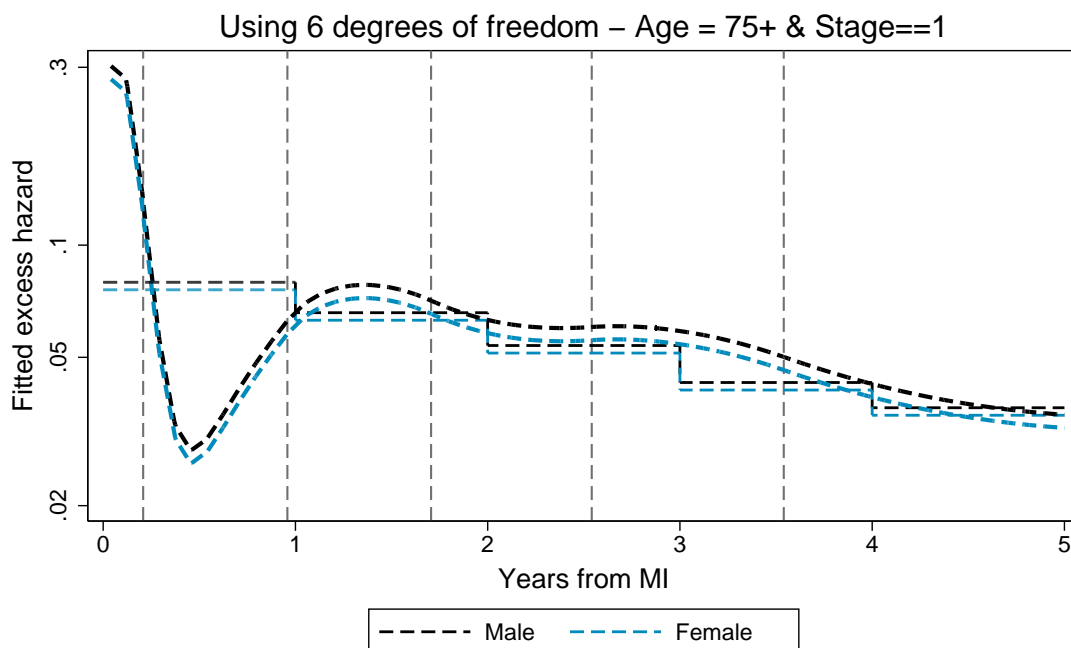
Using 6 degrees of freedom – Age = 75+



## Compare to GLM, Dickman et al. (2004)



## Compare to GLM, Dickman et al. (2004)



## Age continuous

- For a PEH or non-PEH model we could assume age is linear

### Stata output

```
strsrcs agehosp , df(4) bhazard(rate) scale(h)

-----+-----
          |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
xb       |
  agehosp |   .0131148   .0030013    4.37   0.000    .0072323   .0189972
  _cons   |  -3.190079   .2206072   -14.46   0.000   -3.622461  -2.757697
-----+-----
s1       |
  _cons   |   .7594209   .0480404   15.81   0.000    .6652633   .8535784
-----+-----
s2       |
  _cons   |   .1064705   .0297691    3.58   0.000    .0481241   .1648169
-----+-----
s3       |
  _cons   |  - .4495315   .0972077   -4.62   0.000   - .6400551  - .2590078
-----+-----
s4       |
  _cons   |   .564345    .1118058    5.05   0.000    .3452096   .7834805
-----+-----
```

- i.e. for every increase in age by one, the excess hazard rate increases by  $\approx 14\%$

## Age continuous

- Or we could assume that age could also be modelled using splines

```
. rcs agehosp, knots(20 40 60 70 95) gen(agercs)
Variables agercs1 to agercs4 were created
. strsrcs agercs* , df(4) bhazard(rate) scale(h)
```

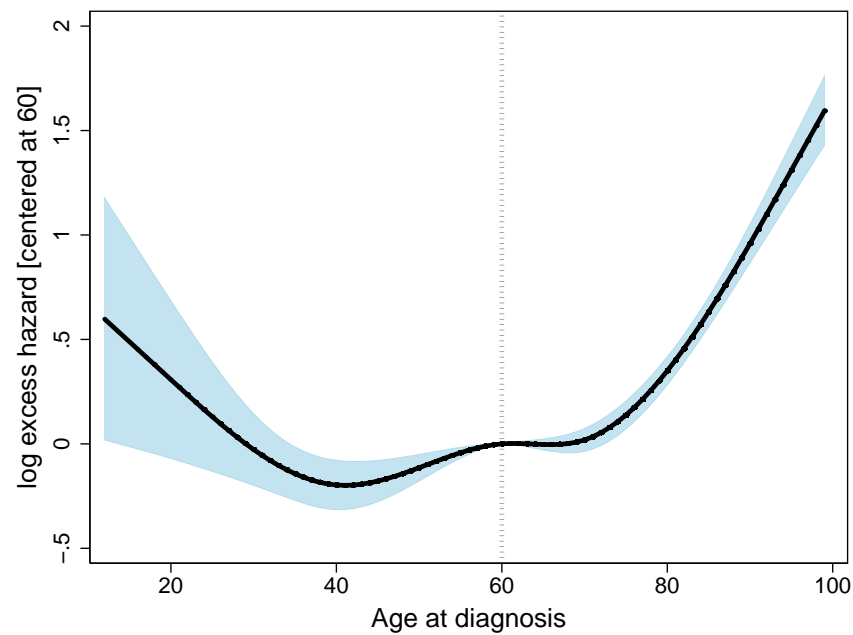
- To get a prediction we can center at age 60 and use predictnl

```
. summ agercs* if agehosp == 60

Variable |      Obs      Mean   Std. Dev.   Min      Max
-----+-----
  agercs1 |       125         60         0         60         60
  agercs2 |       125   -38933.33         0   -38933.33  -38933.33
  agercs3 |       125   -29866.67         0   -29866.67  -29866.67
  agercs4 |       125  -21333.33         0  -21333.33  -21333.33

predictnl xb2 = [xb][agercs1]*(agercs1 - 60) + [xb][agercs2]*(agercs2 + 38933.33) ///
               + [xb][agercs3]*(agercs3 + 29866.67) + [xb][agercs4]*(agercs4 + 21333.33), ci(lo hi)
note: Confidence intervals calculated using Z critical values
```

## Age continuous



## Issues with the method

- Dipping early on - artefact??
- Proportional Odds modeling in relative survival
- Assessing more than one time dependent covariate

# References I

- Bolard, P., Quantin, C., Abrahamowicz, M., Esteve, J., Giorgi, R., Chadha-Boreham, H., Binquet, C., and Faivre, J. (2002). Assessing time-by-covariate interactions in relative survival models using restrictive cubic spline functions. *Journal of Cancer Epidemiology & Prevention*, 7(3):113–122.
- Dickman, P. W., Sloggett, A., Hills, M., and Hakulinen, T. (2004). Regression models for relative survival. *Statistics in Medicine*, 23(1):51–64.
- Giorgi, R., Abrahamowicz, M., Quantin, C., Bolard, P., Esteve, J., Gouvernet, J., and Faivre, J. (2003). A relative survival regression model using b-spline functions to model non-proportional hazards. *Statistics in Medicine*, 22(17):2767–2784.
- Royston, P. and Parmar, M. K. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21(15):2175–2197.