A Comparison of Modeling Scales in Flexible Parametric Models

Noori Akhtar-Danesh, PhD
McMaster University
Hamilton, Canada
daneshn@mcmaster.ca

McMaster University
School of Nursing

Outline

- Background
- A review of splines
- Flexible parametric models
- Results
  - Ovarian cancer
  - Colorectal cancer
- Conclusions

Background

- Cox-regression and parametric survival models are quite common in the analysis of survival data
- Recently, *Flexible Parametric Models* (FPM), have been introduced as an extension to the parametric models such as Weibull model (hazard- scale), loglogistic model (odds-scale), and lognormal model (probit-scale)

Stata Conference 2015, Columbus, USA 3

Objectives & Methods

- In this presentation different FPMs will be compared based on these modeling scales
- Used two subsets of the U.S. National Cancer Institute's Surveillance, Epidemiology and End Results (SEER) dataset from the original 9 registries;
  - Ovarian cancer diagnosed 1991 - 2010
  - Colorectal cancer in men 60^{+} diagnosed 2001 - 2010

Stata Conference 2015, Columbus, USA 4

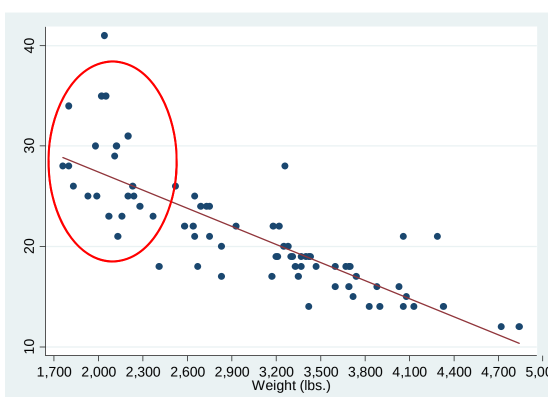
R review of splines

- The daily statistical practice usually involves assessing relationship between one outcome variable and one or more explanatory variables
- We usually assume **linear** relationship between some function of the outcome variable and the explanatory variables
- However, in many situations this assumption may not be appropriate

Linear spline

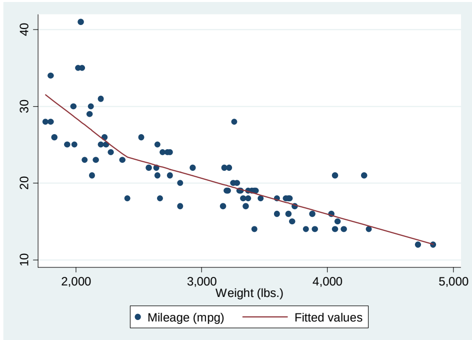
`sysuse auto`

twoway (scatter mpg weight) (lfit mpg weight), xlabel(1700(300)4900)



Linear spline

```
mkspline lnweight1 2400 lnweight2= weight
regress mpg lnweight*
predict linsp
twoway (scatter mpg weight) (line linsp weight, sort)
```



Cubic Splines

- Cubic splines are piecewise cubic polynomials with a separate cubic polynomial fit in each of the predefined number of intervals
- The number of intervals is chosen by the user and the split points are known as *knots*
- *Continuity restrictions* are imposed to join the splines at *knots* to fit a smooth function

Restricted Cubic Splines

- In RCS the spline function is forced (restricted) to be linear before the first and after the last knot (*the boundary knots*)
- When modeling survival time, the boundary knots are usually defined as the minimum and maximum of the uncensored survival times

Stata Conference 2015, Columbus, USA 9

Restricted Cubic Splines

- Let $s(x)$ be the restricted cubic spline function, if we define m interior knots, $k_1,\ldots, k_m$, and two boundary knots, k_{min} and k_{max}, we can write $s(x)$ as a function of parameters γ and some newly defined variables $z_1,\ldots, z_{m+1}$,

$$s(x) = \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + \ldots + \gamma_{m+1} z_{m+1}$$

Stata Conference 2015, Columbus, USA 10

Restricted Cubic Splines

- The derived variables (z_j, also know as the basis functions) are calculated as following

$$\begin{cases} z_1 = x \\ z_j = (x-k_j)_+^3 - \lambda_j (x-k_{\min})_+^3 - (1-\lambda_j)(x-k_{\max})_+^3 \end{cases}$$

where for $j=2,\ldots,m+1$, and

$(x-k_j)_+^3 = (x-k_j)^3$ if it is positive and 0 otherwise

$$\lambda_j = \frac{k_{\max} - k_j}{k_{\max} - k_{\min}}$$

(Royston & Lambert, 2011)

Stata Conference 2015, Columbus, USA 11

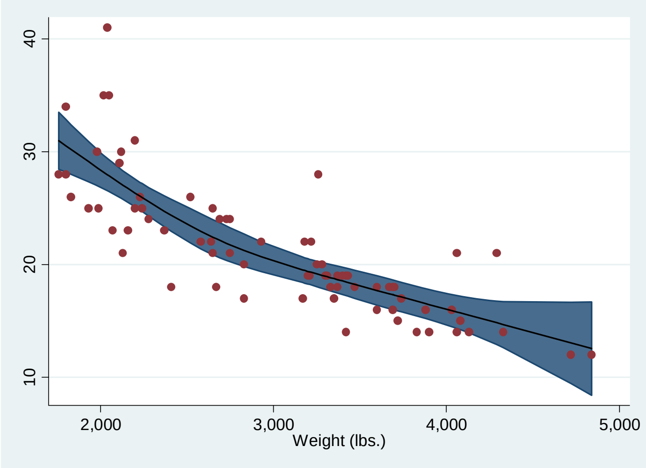
Restricted Cubic Splines

- These RCSs can be calculated using a number of **`Stata`** commands, including `mkspline` (an official **`Stata`** command), `rcsgen`, and `splinegen` (two user written commands)
- The `rcsgen` command can orthogonalize the derived spline variables which can lead to more stable parameter estimates and quicker model convergence

Stata Conference 2015, Columbus, USA 12

Flexible Parametric Models

rcsgen; *an alternative spline macro*



Stata Conference 2015, Columbus, USA 13

FPM: *Royston-Parmer (RP) Models*

- RP models are a extension of the parametric models (Weibull, log-logistic, and log-normal) which offer greater flexibility with respect to shape of the survival distribution
- The additional flexibility of an RP model is because, for instance for a hazard model, it represents the **baseline distribution function** as a restricted cubic spline function of log time instead of simply as a linear function of log time
- The complexity of modeling spline functions is determined by the number and positions of the knots in the log time

Stata Conference 2015, Columbus, USA 14

FPM: *Royston-Parmer (RP) Models*

- Spline models can be chosen by the appearance of the survival functions, hazard functions, etc. or more formally, by minimizing the value of an information criterion [Akaike (AIC) or Bayes (BIC)]
- Estimation of parameters is by maximum likelihood

Stata Conference 2015, Columbus, USA 15

FPM: *A review of Weibull distribution*

- The cumulative hazard function for a Weibull distribution is

$$H(t) = \lambda t^{p}$$

- To make it consistent with rest of this presentation let's change the notation as

$$H(t) = \lambda t^{\gamma_1}$$

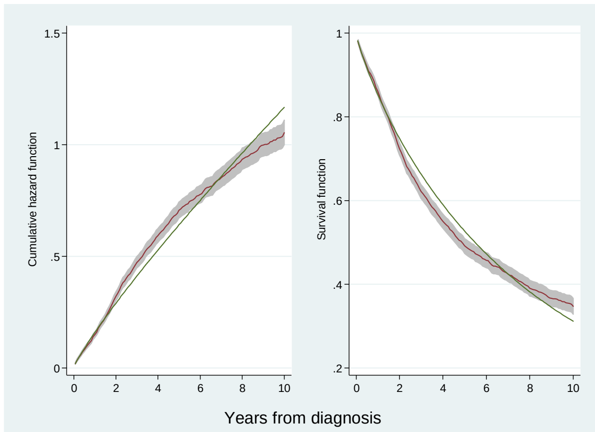
where γ_1 is the shape parameter. Then, the Weibull hazard function is

$$h(t) = dH(t)/dt = \lambda \gamma_1 t^{\gamma_1 - 1}$$

Stata Conference 2015, Columbus, USA 16

FPM: *A review of Weibull distribution*

- Women age 60-69 diagnosed with ovarian cancer



FPM: *A review of Weibull distribution*

- One reason that a Weibull model does not fit very well to the dataset is that it has a monotonic hazard function
- To have a more flexible form, we begin by writing the Weibull cumulative hazard function in logarithmic form

$$\ln H(t) = \ln \lambda + \gamma_1 \ln t = \gamma_0 + \gamma_1 \ln t$$

- Now, suppose that $f(t;\gamma)$ represents some general family of nonlinear functions of time t, with some parameter vector γ and

$$\ln H(t) = f(t;\gamma)$$

FPM: *Royston-Parmer (RP) Models*

- Because cumulative hazard functions are monotonic in time, $f(t;\gamma)$ must be monotonic too
- Two potentially appropriate functions are fractional polynomials (Royston & Altman 1994) and splines (de Boor 2001)

Stata Conference 2015, Columbus, USA 19

FPM: *Royston-Parmer (RP) Models*

- We write a restricted cubic spline function as $s(\ln t;\gamma)$ instead of $f(t;\gamma)$ with s standing for spline and $\ln t$ to emphasize that we are working on the scale of log time

$$\ln H(t) = s(\ln t;\gamma) = \gamma_0 + \gamma_1 \ln t + \gamma_2 z_1(\ln t) + \gamma_3 z_2(\ln t) + \dots$$

where $\ln t$, $z_1(\ln t)$, $z_2(\ln t)$, ..., are the basis functions of the restricted cubic spline

Stata Conference 2015, Columbus, USA 20

FPM: *Royston-Parmer (RP) Models*

- When we specify one or more knots, the spline function includes a constant term (γ_0), a linear function of $\ln t$ with parameter γ_1, and a basis function for each knot
- By convention, the "no knots" case for a hazard model corresponds to the linear function, $s(\ln t;\gamma)= \gamma_0+ \gamma_1 \ln t$, which is the Weibull model

Stata Conference 2015, Columbus, USA 21

FPM: *Royston-Parmer (RP) Models*

- We estimate the γ parameters by maximum likelihood method using the **`stpm2`** routine (Lambert & Royston 2009)
- We identify **`df`** for each model based on AIC criteria and evaluate the variables in the model using **`lrtest`**
- We use options of **`hazard, odds, and normal`** in **`stpm2`** for fitting different scales

Stata Conference 2015, Columbus, USA 22

FPM: Ovarian cancer

```
. tab agegrp
           Age group |      Freq.     Percent        Cum.
        ------------+-----------------------------------
        40- 49 years |      2,700       19.55       19.55
        50- 59 years |      3,896       28.21       47.76
        60- 69 years |      3,466       25.10       72.86
        70- 79 years |      2,606       18.87       91.73
          >=80 years |      1,142        8.27      100.00
        ------------+-----------------------------------
               Total |     13,810      100.00
gen year=DATE_yr-1990
mkspline yearsp=year, cubic nknots(3)

stpm2 agegrp2-agegrp5 yearsp*, df(7) tvc(agegrp2- ///
      agegrp5 yearsp1) dftvc(2) ///
      scale(hazard) eform nolog
```

Stata Conference 2015, Columbus, USA 23

FPM: Ovarian cancer

```
. estat ic

        Scale     |      AIC
   -------------+-------------
       Hazard     |  35615.92
       Odds       |  35616.51
       Normal     |  35564.12
```

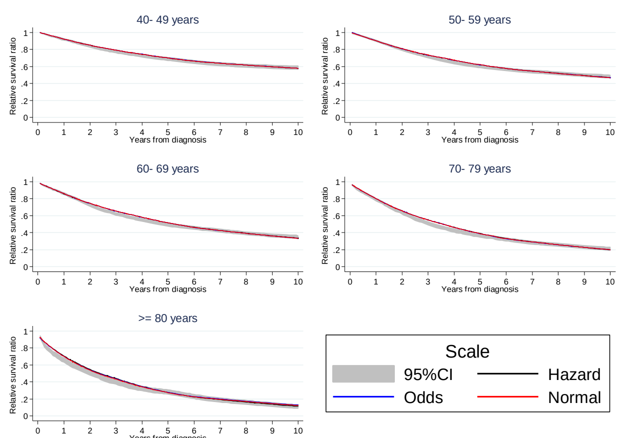
Stata Conference 2015, Columbus, USA 24

FPM: Ovarian cancer

```
predict s1, survival

gen time1=1

predict surv1lyr, survival timevar(time1) //ci

gen time5=5

predict surv15yr , survival timevar(time5) //ci


range tempt2 0.05 5 2000

forvalues i=1/5 {

                predict haz1`i', hazard per(100) timevar(tempt2) ///
                        at(agegrp`i' 1 yearsp1 15 yearsp2 6.508929) zeros
           }
```

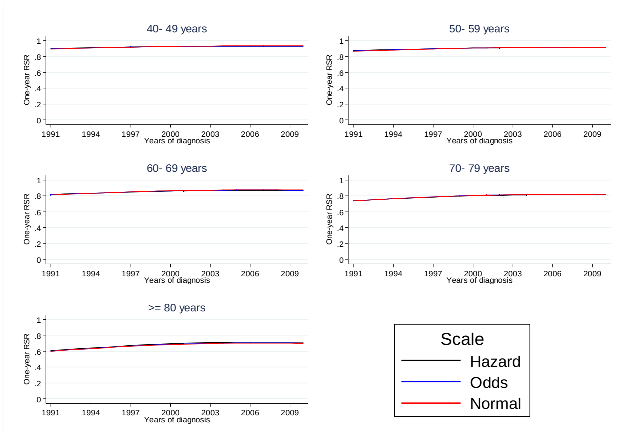
Stata Conference 2015, Columbus, USA 25

FPM: Ovarian cancer

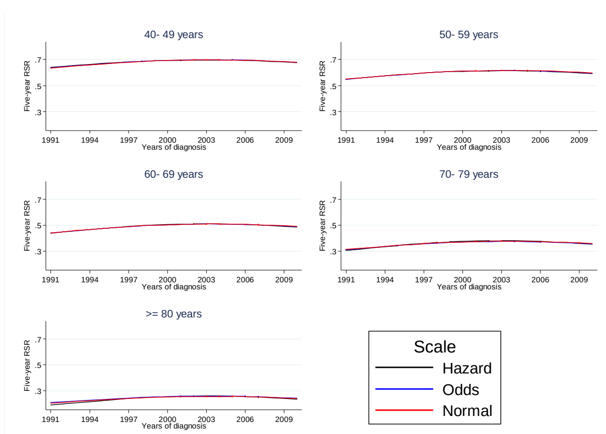


Stata Conference 2015, Columbus, USA 26

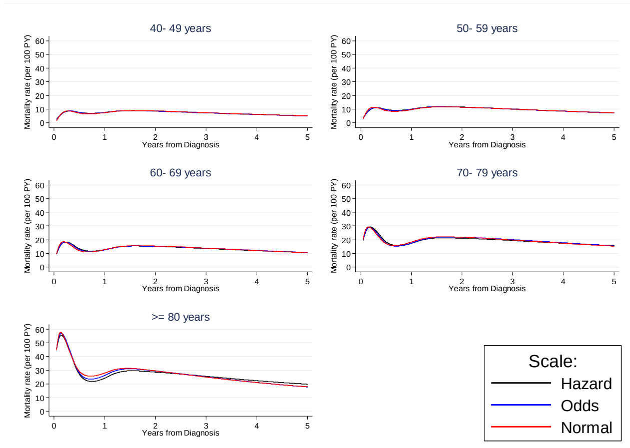
FPM: Ovarian cancer



FPM: Ovarian cancer



FPM: Ovarian cancer



FPM: Colorectal cancer

```
. tab agegrp

    Age group    |      Freq.     Percent        Cum.
    -------------+-----------------------------------
    60- 69 years |     14,837       35.32       35.32
    70- 79 years |     16,026       38.16       73.48
      >=80 years |     11,139       26.52      100.00
    -------------+-----------------------------------
           Total |     42,002      100.00
```

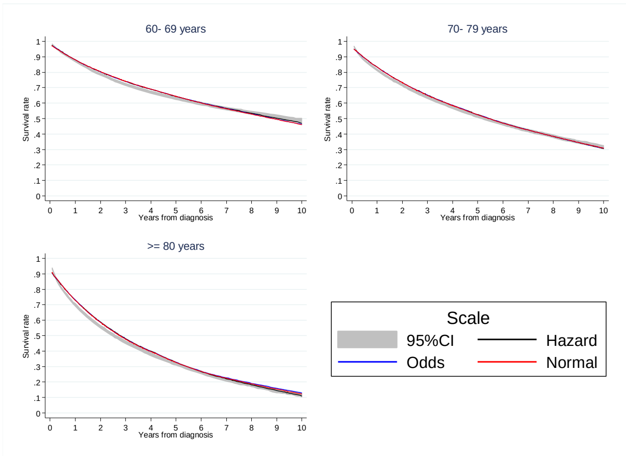
```
stpm2 agegrp2-agegrp3 year, df(9) tvc(agegrp2- ///
   agegrp3) dftvc(2) scale(hazard) eform nolog
```

FPM: Colorectal cancer

```
. estat ic

        Scale      |      AIC
  -------------+-------------
        Hazard     | 111997.6
        Odds       | 112056.2
        Normal     | 111829.9
```

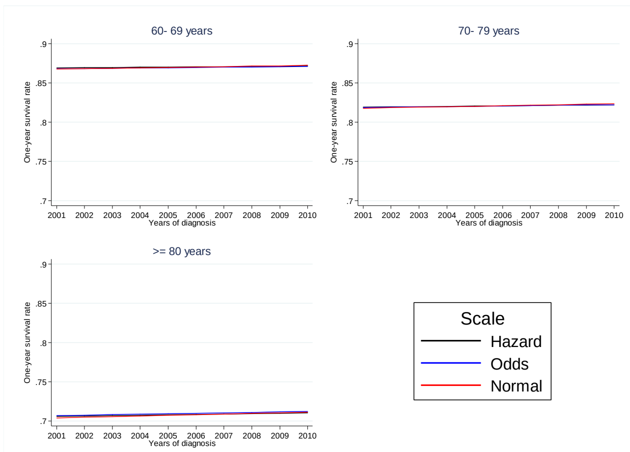
Stata Conference 2015, Columbus, USA 31

FPM: Colorectal cancer



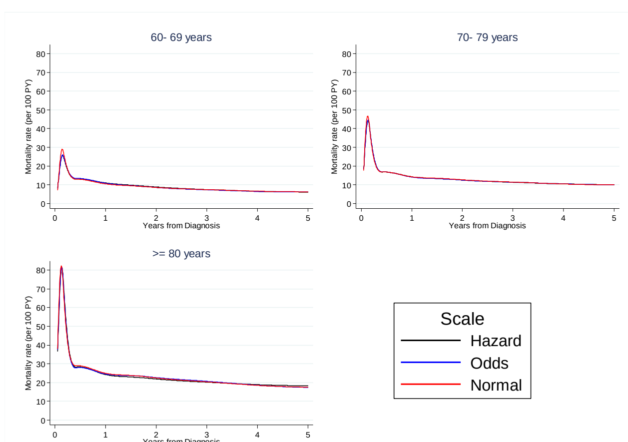
Stata Conference 2015, Columbus, USA 32

FPM: Colorectal cancer



Stata Conference 2015, Columbus, USA 33

FPM: Colorectal cancer



Stata Conference 2015, Columbus, USA 34

Conclusion

- In general, there were no substantial differences between the estimates from the three modeling scales, although the probit-scale showed slightly better fit based on the Akaike information criterion (AIC) for both datasets

References

- de Boor, C. 2001. *A Practical Guide to Splines*, Revised ed ed. New York, Springer.
- Durrleman, S. & Simon, R. 1989. Flexible regression models with cubic splines. *Stat.Med.*, 8, (5) 551-561 available from: PM:2657958
- Lambert, P.C. & Royston, P. 2009. Further development of flexible parametric models for survival analysis. *The Stata Journal*, 9, 265-290
- Royston, P. & Altman, D.G. 1994. Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling (with discussion). *Applied Statistics*, 43, 429-467
- Royston, P. & Lambert, P.C. 2011. *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model* Stata Press.
- Royston, P. & 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. *Stat.Med.*, 21, (15) 2175-2197 available from: PM:12210632