rstpm2: a simple guide

Mark Clements
Karolinska Institutet

Abstract

This vignette provides a simple guide to flexible parametric models provided by **rstpm2**.

Keywords: survival, splines.

1. Introduction

The **rstpm2** package supports *flexible parametric survival models* to model *time-to-event* data. These models are fully parametric for the survival function. These models are particularly useful for:

- Estimating predictions for hazards, hazard differences, hazard ratios, survival, survival differences and survival ratios, restricted mean survival
- Estimating marginal predictions, including standardised survival and standardised survival differences
- Modelling time-varying effects, including time-varying hazards ratios.

This guide is intended to provide an accessible guide to some of the models and predictions provided by flexible parametric survival models. This guide can then be followed by the other vignette, which provides a more complete mathematical presentation.

For this guide, we describe the most common flexible parametric survival model, which is a *proportional hazards* model. Let the survival function $S(t|\boldsymbol{x}) = \Pr(T > t|\boldsymbol{x})$ for random variable T at time t and covariates $\boldsymbol{x} = (x_j)$ be modelled by

$$S(t|\boldsymbol{x}) = \exp\left(-\exp\left(s(\log(t);\boldsymbol{\gamma}) + \sum_j \beta_j x_j\right)\right)$$

for some parametric smooth function $s(u;\boldsymbol{\gamma})$, for parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, and for j being an index over the covariates. For this model, we use a smooth function to model the baseline log cumulative hazard function and include a linear predictor to model the covariates. To see that this is a proportional hazards model, we can see that the cumulative hazard and hazard functions are, respectively,

$$H(t|\boldsymbol{x}) = -\log(S(t|\boldsymbol{x})) = \exp\left(s(\log(t);\boldsymbol{\gamma}) + \sum_j \beta_j x_j\right)$$

$$h(t|\boldsymbol{x}) = \frac{\mathrm{d}}{\mathrm{d}t}H(t|\boldsymbol{x}) = \exp\left(s(\log(t);\boldsymbol{\gamma}) + \sum_j \beta_j x_j\right) \times \frac{\mathrm{d}\, s(\log(t);\boldsymbol{\gamma})}{\mathrm{d}t}$$

Now, for two sets of covariates $\boldsymbol{x}_1 = (x_{1j})$ and $\boldsymbol{x}_2 = (x_{2j})$, we have the hazard ratio

$$\begin{aligned}\frac{h(t|\boldsymbol{x}_2)}{h(t|\boldsymbol{x}_1)} &= \frac{\exp(s(\log(t);\boldsymbol{\gamma}) + \sum_j \beta_j x_{2j}) \times \frac{\mathrm{d}s(\log(t);\boldsymbol{\gamma})}{\mathrm{d}t}}{\exp(s(\log(t);\boldsymbol{\gamma}) + \sum_j \beta_j x_{1j}) \times \frac{\mathrm{d}s(\log(t);\boldsymbol{\gamma})}{\mathrm{d}t}} \\ &= \exp\left(\sum_j \beta_j (x_{2j} - x_{1j})\right)\end{aligned}$$

If the covariates only vary by one for the jth covariate, such that $x_{2j} = x_{1j} + 1$ and $x_{2j'} = x_{1j'}$ for $j' \neq j$, then the hazard ratio is equal to $\exp(\beta_j)$ for all t and for all values of the other covariates.

We can motivate this model as an extension of exponential (or Poisson) regression. If we assume that the rates are constant over time and proportional with respect to covariates, then we have an exponential distribution with a hazard $h(t|\boldsymbol{x}) = \exp(\gamma_0 + \sum_j \beta_j x_j)$ for log baseline hazard γ_0, with a survival function $S(t|\boldsymbol{x}) = \exp(-\exp(\gamma_0 + \log(t) + \sum_j \beta_j x_j))$. The flexible parametric survival models generalise the function $\gamma_0 + \log(t)$ to some smooth function $s(\log(t);\boldsymbol{\gamma})$.

The default smoother provided by `rstpm2::stpm2` is a natural spline, such that

$$s(\log(t);\boldsymbol{\gamma}) = \sum_{k=1}^{K} B_k(\log(t))\gamma_k$$

where $B_k(\log(t))$ is a natural spline basis with K degrees of freedom. Natural splines have the property that the function is cubic between internal *knots* (fixed points that default to quantiles of the event times) and linear outside of the knot boundaries, with continuous derivatives at the knots. Heuristically, splines provide a flexible functional form that looks "nice". The basis can be defined in several ways (e.g. using a truncated power basis as used in Stata), while we use the approach used by the `splines::ns` function, which uses a matrix projection of the second derivatives at the knot boundaries.

We fit this model using *maximum likelihood estimation* for right censored and left truncated data. Variance estimation assumes that the parameters are asymptotically normal, with variable for predictions calculated using the *multivariate delta method*.

2. An example

We begin with some simple proportional hazard models using the `brcancer` dataset. We first fit a Cox regression with a single indicator for whether an a breast cancer patient was randomised to hormonal treatment. From the output, we see that hormonal treatment is associated with improved survival (HR=0.69, 95% CI: 0.54, 0.89).

```
> library(survival)
> library(rstpm2)
> brcancer <- transform(brcancer, recyear=rectime / 365.24)
> fit.cox <- coxph(Surv(recyear,censrec==1)~hormon, data=brcancer)
> summary(fit.cox)

Call:
coxph(formula = Surv(recyear, censrec == 1) ~ hormon, data = brcancer)

  n= 686, number of events= 299
```

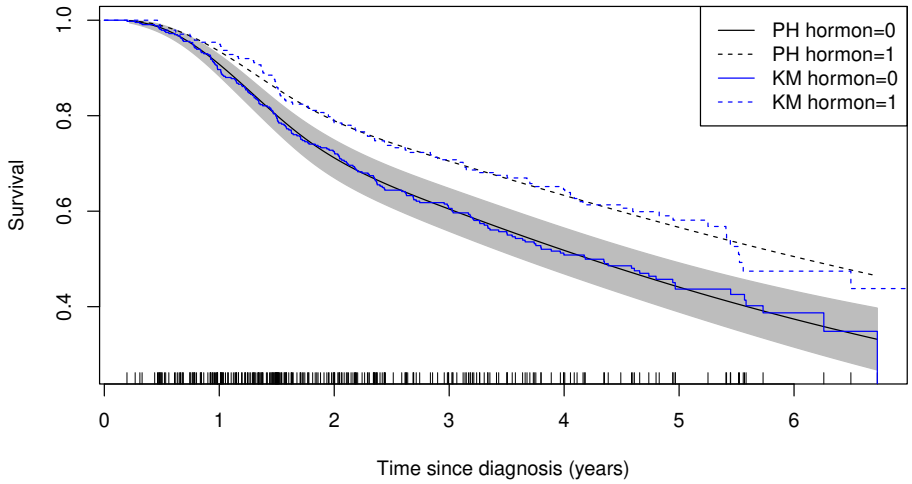
```
          coef exp(coef) se(coef)      z Pr(>|z|)
hormon -0.3640    0.6949   0.1250 -2.911   0.0036 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

       exp(coef) exp(-coef) lower .95 upper .95
hormon    0.6949      1.439    0.5438    0.8879

Concordance= 0.543  (se = 0.014 )
Likelihood ratio test= 8.82  on 1 df,   p=0.003
Wald test            = 8.47  on 1 df,   p=0.004
Score (logrank) test = 8.57  on 1 df,   p=0.003
```

We can fit a flexible parametric survival model with `rstpm2::stpm2` using very similar syntax, with an additional argument `df=4` to specify four degrees of freedom for the baseline smoother (typical values for the degrees of freedom are 2–6). From the output, the model parameters include an intercept term, time-invariant log-hazard ratios, and parameters for the baseline smoother. The hazard ratio for hormonal treatment is 0.69 (95% CI: 0.56, 0.84), which is a similar point estimate and a more narrow confidence interval than Cox regression.

```
> fit <- stpm2(Surv(recyear,censrec==1)~hormon, data=brcancer, df=4)
> summary(fit)

Maximum likelihood estimation

Call:
stpm2(formula = Surv(recyear, censrec == 1) ~ hormon, data = brcancer,
    df = 4)

Coefficients:
                           Estimate Std. Error z value     Pr(z)
(Intercept)                -6.79773    0.72642 -9.3578 < 2.2e-16 ***
hormon                     -0.36406    0.12491 -2.9144  0.003563 **
nsx(log(recyear), df = 4)1  5.69995    0.71677  7.9523 1.830e-15 ***
nsx(log(recyear), df = 4)2  4.85614    0.48002 10.1166 < 2.2e-16 ***
nsx(log(recyear), df = 4)3 10.13327    1.41267  7.1731 7.331e-13 ***
nsx(log(recyear), df = 4)4  4.70626    0.33016 14.2545 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

-2 log L: 1684.412

> eform(fit)[2,]

exp(beta)     2.5 %    97.5 %
0.6948520 0.5636727 0.8449352
```

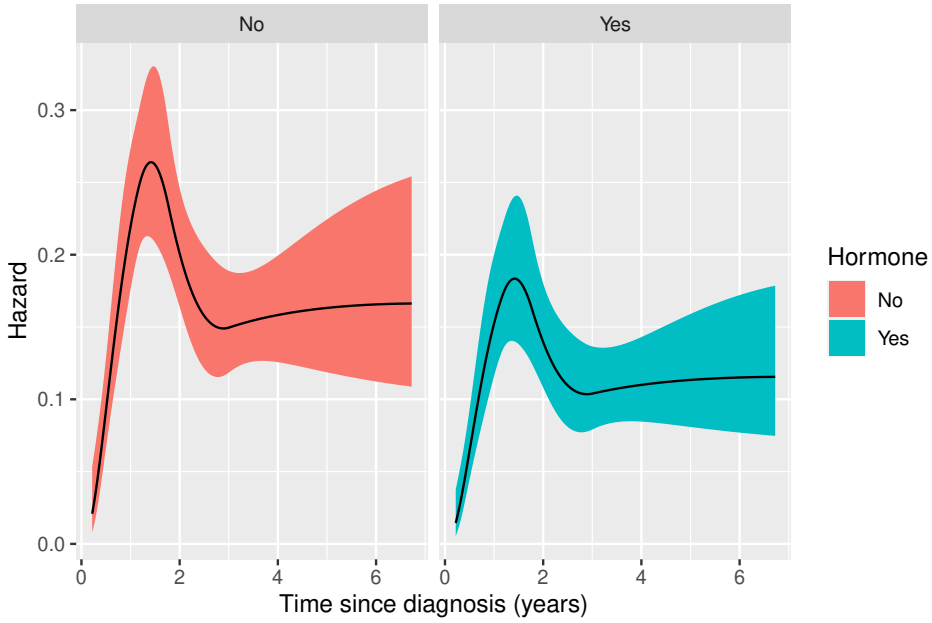
The flexible parametric survival models can be used to estimate a variety of parameters. For example, we can easily estimate survival and compare with predictions with the non-parametric Kaplan-Meier curves. From the output, we note that...

```
> plot(fit, newdata=data.frame(hormon=0), xlab="Time since diagnosis (years)")
> lines(fit, newdata=data.frame(hormon=1), lty=2)
> lines(survfit(Surv(recyear,censrec==1)~hormon, data=brcancer), col="blue", lty=1:2)
> legend("topright", c("PH hormon=0","PH hormon=1","KM hormon=0","KM hormon=1"),
+        lty=1:2, col=c("black","black","blue","blue"))
```

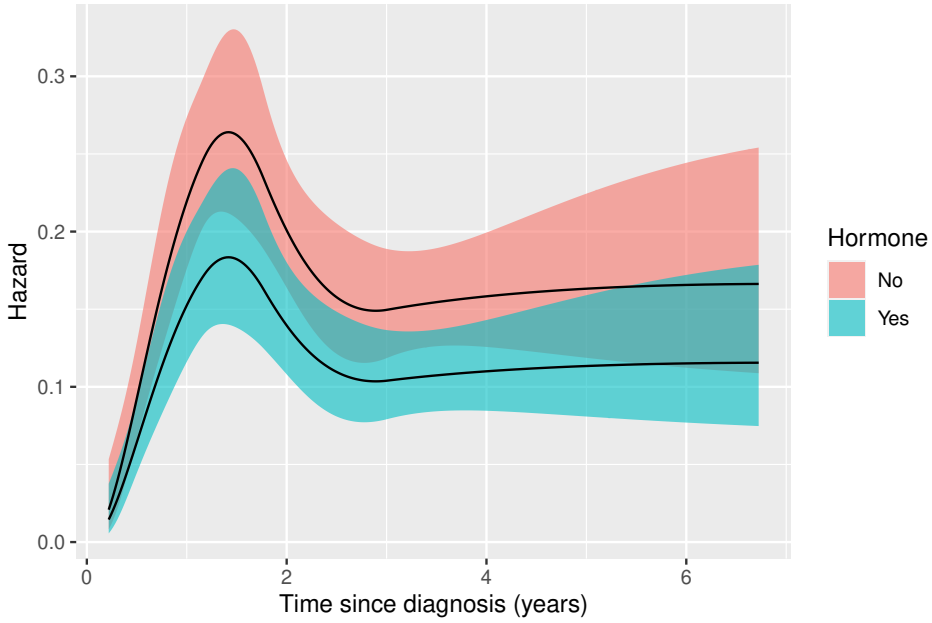


We can also plot the hazards using **ggplot2**. This requires that we predict using `grid=TRUE` to get a time grid, with `full=TRUE` to include the covariates from `newdata`, and with `se.fit=TRUE` to get the confidence intervals.

```
> library(ggplot2)
> predHormon <- predict(fit, newdata=data.frame(hormon=0:1),
+                       type="hazard", grid=TRUE, full=TRUE, se.fit=TRUE)
> predHormon <- transform(predHormon,Hormone=factor(hormon,labels=c("No","Yes")))
> ggplot(predHormon,
+        aes(x=recyear,y=Estimate,ymin=lower,ymax=upper,fill=Hormone)) +
+     facet_grid(~Hormone) +
+     xlab("Time since diagnosis (years)") +
+     ylab("Hazard") +
+     geom_ribbon() +
+     geom_line()
```



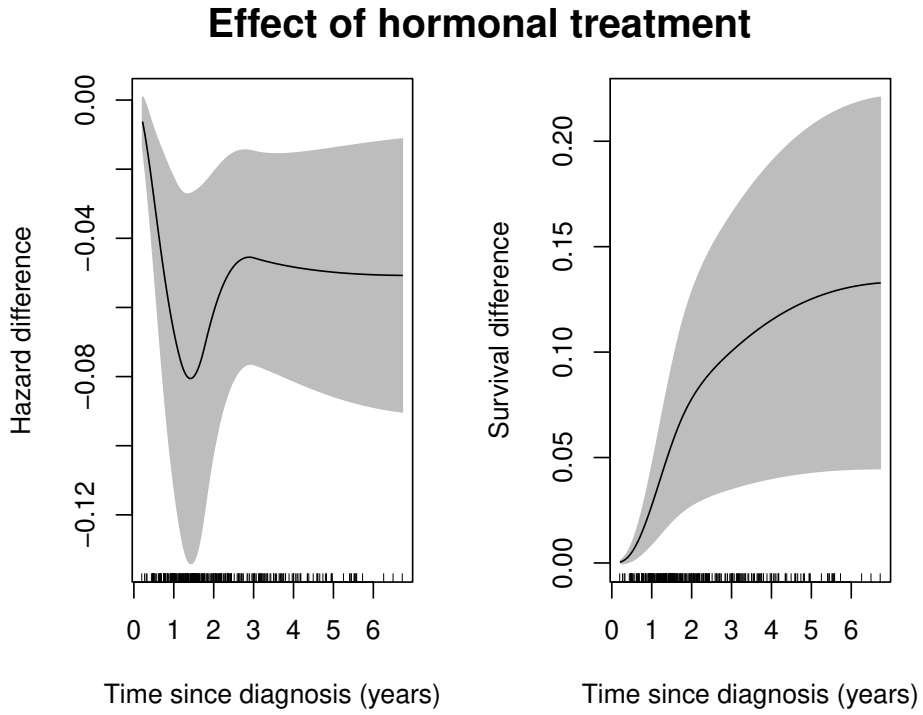

```
> ggplot(predHormon,
+        aes(x=recyear,y=Estimate,ymin=lower,ymax=upper,fill=Hormone)) +
+     xlab("Time since diagnosis (years)") +
+     ylab("Hazard") +
+     geom_ribbon(alpha=0.6) +
+     geom_line()
```



Usefully, we can also estimate survival differences and hazard differences. We define the survival differences using a reference covariate pattern using the newdata argument, and then define an exposed function which takes the newdata and transforms for the 'exposed' covariate pattern. As an example:

```
> par(mfrow=1:2)
> plot(fit,newdata=data.frame(hormon=0), type="hdiff",
```

```
+      exposed=function(data) transform(data, hormon=1),
+      xlab="Time since diagnosis (years)")
> plot(fit,newdata=data.frame(hormon=0), type="sdiff",
+      var="hormon",
+      xlab="Time since diagnosis (years)")
> mtext("Effect of hormonal treatment", outer = TRUE, line=-3, cex=1.5, font=2)
```



Affiliation:

Mark Clements
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
Email: mark.clements@ki.se