flexsurv: flexible parametric survival modelling in R

Christopher H. Jackson
MRC Biostatistics Unit, Cambridge, UK
chris.jackson@mrc-bsu.cam.ac.uk

Abstract

flexsurv is an R package for fully-parametric modelling of survival data. Any parametric time-to-event distribution may be fitted if the user supplies at minimum a probability density or hazard function. Many standard survival distributions are built in, and also the three and four-parameter generalized gamma and F models. Any parameter of the distribution can be modelled as a linear or log-linear function of covariates. Another built-in model is the spline model of Royston and Parmar, in which both baseline survival and covariate effects can be arbitrarily flexible parametric functions of time.

flexsurv is intended to be similar to **survival**: right-censoring or left-truncation are specified in `Surv` objects, and the main model-fitting function, `flexsurvreg`, uses the familiar syntax of `survreg`. **flexsurv** also enhances the **mstate** package (Putter et al) by providing cumulative incidences for fully-parametric multi-state models.

Keywords:˜—!!!—at least one keyword is required—!!!—.

1. Package motivation and design

adv of parametric over cox. examples in HE ref stata based on survival. right cens, left trunc,

The `survreg` function in **survival** only supports two-parameter (location/scale) distributions, though users can supply their own distributions.

Stata has a nice `streg`. the `stpm2` spline model More generally `stgenreg` has general. ours can avoid num integ The `flexsurv` is similar in spirit

review other packages in survival view: many model-specific packages

2. Generic parametric survival models

2.1. Definitions

The general model that **flexsurv** fits has probability density function

$$f(t|\mu(\mathbf{z}), \boldsymbol{\alpha}(\mathbf{z})), \quad t \geq 0 \tag{1}$$

$\mu = \alpha_0$ is the parameter of primary interest, which usually governs the mean or location of the distribution. Other parameters $\alpha = \alpha_1, \ldots, \alpha_R$ are called "ancillary" and determine the variance, shape or higher moments. All parameters may depend on a vector of covariates $\mathbf{z}$

through link-transformed linear models $g_0(\mu) = \boldsymbol{\gamma}_0'\mathbf{z}$ and $g_r(\alpha_r) = \boldsymbol{\gamma}_r'\mathbf{z}$. $g(a)$ will typically be $\log(a)$ if a is defined to be positive, or $g(a) = a$ if a is unrestricted.

We also define (suppressing the conditioning for clarity) the cumulative distribution function $F(t)$, survivor function $S(t) = 1 - F(t)$, cumulative hazard $H(t) = -\log S(t)$ and hazard $h(t) = f(t)/S(t)$.

Let $t_i : i = 1, \ldots n$ be a sample of times from individuals i. Let $c_i = 1$ if t_i is an observed death time, or $c_i = 0$ if t_i is a right-censoring time, thus the true death time is known only to be greater than t_i. Also let s_i be corresponding left-truncation times, meaning that individual i is only observed conditionally on survival up to s_i, thus $s_i = 0$ if there is no left-truncation.

The likelihood for the parameters in model (1), given the corresponding data vectors, is

$$l(\mu, \boldsymbol{\alpha}|\mathbf{t}, \mathbf{c}, \mathbf{s}) = \left\{ \prod_{i:\ c_i=1} f_i(t_i) \prod_{i:\ c_i=0} S_i(t_i) \right\} / \prod_i S_i(s_i)$$

EXAMPLE DATASET HERE. bc? not oral or franc. ask unit, rita3? latimer?

2.2. Other software

2.3. Model fitting syntax

The main model-fitting function is called `flexsurvreg`. Its first argument is an R *formula* object. The left hand side of the formula is defined by a `Surv` function from the **survival** package. This indicates here that the response variable is `recyrs`, and that these are observed death and censoring times when the variable `censrec` is 1 or 0 respectively. If we also had left-truncation times in a variable called `start`, the response would be `Surv(start,recyrs,censrec)`. All of these variables are in the data frame called `bc`.

In order to fit a model, needs to know at least its probability density function.

```
library(flexsurv)
flexsurvreg(Surv(recyrs, censrec) ~ group, data=bc, dist="weibull")

## Warning:  NaNs produced
## Warning:  NaNs produced
## Warning:  NaNs produced

##
## Call:
## flexsurvreg(formula = Surv(recyrs, censrec) ~ group, data = bc,     dist = "weibull")
##
## Estimates:
##              data mean  est      L95%     U95%     se       exp(est)
## shape             NA     1.3797   1.2548   1.5170   0.0668       NA
## scale             NA    11.4229   9.1818  14.2110   1.2728       NA
## groupMedium   0.3338    -0.6136  -0.8623  -0.3649   0.1269   0.5414
## groupPoor     0.3324    -1.2122  -1.4583  -0.9661   0.1256   0.2975
```

	Parameters	Density function
Exponential	rate	dexp
Weibull	shape, scale	dweibull
Gamma	shape, rate	dgamma
Log-normal	meanlog, sdlog	dlnorm
Gompertz	shape, rate	dgompertz
Generalized gamma (Prentice)		dgengamma
Generalized gamma (Stacy 1962)		dgengamma.orig
Generalized F (Prentice)		dgenf
Generalized F		dgenf.orig

Table 1: Built-in parametric survival distributions in **flexsurv**

```
##                L95%     U95%
## shape             NA       NA
## scale             NA       NA
## groupMedium   0.4222   0.6943
## groupPoor     0.2326   0.3806
##
## N = 686,  Events: 299,  Censored: 387
## Total time at risk: 2113
## Log-likelihood = -811.9, df = 4
## AIC = 1632
```

2.4. Built-in survival distributions

If the argument `dist` is a string, this denotes a built-in survival distribution. The built-in distributions are listed in Table 1. In each case the parameterisation of the distribution and the probability density is defined by a function of the same name as codedist, but preceded with `d`, for example if `dist="weibull"`, the density function is `dweibull`. For the exponential (`dexp`), Weibull, gamma (`dgamma`) and log-normal (`dlnorm`), these are provided with standard R installations. For all other models these are supplied in **flexsurv**

In addition **flexsurv** defines the generalized gamma and generalized F distributions, each in two parameterisations Cox, Chu, Schneider, and Muñoz (2007) Prentice (1974) **?** Cox (2008) Prentice (1975) and a Gompertz distribution useful as reduce stable versions

For all built-in distributions, **flexsurv** also defines functions beginning h or H for the hazard or cumulative hazard.

2.5. Supplying own distributions

Suppose we know the distribution

Many contributed R packages contain density and cumulative distribution functions for positive distributions (ref eha, VGAM, ActuDistns)

User-specified p function, d function, hazard since loads of contributed distributions fixed-dimension. In Section 3.1

Distribution exists in another package, but may be parameterised

Example: Gompertz-Makeham

optimisation methods, derivatives , parallel processing with pnmath

Demo on at least one dataset: stgenreg uses bc example i think

test if we can do the gen gamma prop haz trick in stgenreg paper

Basic Weibull prop haz model in stgenreg. Advantage is that it's just as fast as R built in stuff.

2.6. Output functions

`summary.flexsurvreg` calculates the estimated survival, hazard or cumulative hazard at a series of times and for specified covariate values. Confidence intervals are produced by simulating a large sample from the asymptotic normal distribution of the maximum likelihood estimates γ OR WHATEVER, via the function `normboot.flexsurvreg`. The default `plot` method for `flexsurvreg` objects graphs these fitted trajectories against non-parametric estimates based on Kaplan-Meier or kernel estimation (REF muhaz), while the `lines` method adds lines to an existing plot. REFER TO EXAMPLE FIGURE

Any user-defined function of the basic model parameters γ OR WHATEVER and time can also be summarised in the same way. For example, in a non-proportional hazards model, the hazard ratio between two groups of interest varies through time. To plot this trajectory, and confidence intervals. EXAMPLE FROM SPLINE.

Restricted mean survival: say of interest. ref royston + parmar

3. Spline models

parameters are vectors, different design

relation to fractional polynomials (see **mfp** for continuous covariates, slightly diff)

stgenreg has demo of spline modelling on the log hazard scale. Can we do this using a generic distribution? (advantage: when there are multiple time dependent effects, the interpretation of the time-dependent hazard ratios is simplified as they do not depend on values of other covariates, which is the case when modelling on the cumulative hazard scale (Royston and Lambert 2011).

Demo on at least one dataset

3.1. General-dimension models

The spline model above is an example of a model where the general parametric form can be written explicitly as in model 1, but the length of alpha is arbitrary. semi-parametric

flexsurv has the tools to deal with any

where are vectors

4. Multi-state models

enhances mstate

5. Potential extensions

relative survival interval censoring frailty what else does survival do

A. Acknowledgements

Thanks to Milan Bouchet-Valat.

References

Cox C (2008). "The generalized F distribution: An umbrella for parametric survival analysis." **27**, 4301–4312.

Cox C, Chu H, Schneider MF, Muñoz A (2007). "Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution." **26**, 4352–4374.

Prentice RL (1974). "A log gamma model and its maximum likelihood estimation." *Biometrika*, **61**(3), 539–544.

Prentice RL (1975). "Discrimination among some parametric models." *Biometrika*, **62**(3), 607–614.