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.

Any output function

The main model-fitting function, `flexsurvreg`, uses the familiar syntax of `survreg` from the standard **survival** package — censoring or left-truncation are specified in `Surv` objects. **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. Motivation and design

The Cox model is ubiquitous in medical research, since the effects of predictors of survival can be estimated without needing to supply a baseline survival distribution that might be wrong. However, fully-parametric models have many advantages, and even the originator of the Cox model has expressed a preference for parametric modelling (Reid 1994). Fully-specified models help to understand the change in hazard through time, and help with prediction and extrapolation. For example, the mean survival $E(T) = \int_0^\infty S(t)$, used in health economic evaluations (Latimer 2013), needs the survivor function $S(t)$ to be fully-specified for all times t.

flexsurv allows parametric distributions of arbitrary complexity to be fitted to survival data, gaining the convenience of parametric modelling, while avoiding the risk of model misspecification. Built-in choices include splines with any number of knots (Royston and Parmar 2002) and 3–4 parameter generalized gamma and F distribution families. Any user-defined model may be employed by supplying at minimum an R function to compute the probability density or hazard, and ideally also its cumulative form. Any parameters may be modelled in terms of covariates, and any function of the parameters may be printed or plotted in model summaries.

flexsurv is intended as a general platform for survival modelling in R. It is similar in spirit to the Stata packages **stpm2** (Lambert and Royston 2009) for spline-based survival modelling, and **stgenreg** (Crowther and Lambert 2013) for fitting survival models with user-defined hazard functions using numerical integration. The `survreg` function in the R package **survival** only supports two-parameter (location/scale) distributions, though users can supply their own distributions if they can be parameterised in this form. Many other contributed R packages can fit survival models, e.g. **eha** (BrostrÃűm 2014), **VGAM** (Yee and Wild 1996), though these are either limited to specific distribution families not specifically designed for survival analysis, or (**ActuDistns**, Nadarajah and Bakar (2013)) contain only the definitions of distribution functions. **flexsurv** enables distribution functions provided by such packages to be employed in models. An advantage over **stgenreg** is that numerical integration can be avoided if the analytic cumulative distribution or hazard can be supplied, and optimisation can also be speeded by supplying analytic derivatives. **flexsurv** also has features for multi-state modelling and interval censoring, and general output reporting. It employs functional programming to work with user-defined or existing R functions.

2. General parametric survival model

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 $\boldsymbol{\alpha} = \alpha_1, \ldots, \alpha_R$ are called "ancillary" and determine the shape, variance 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 $g_r(\alpha_r) = \boldsymbol{\gamma}_r^{'}\mathbf{z}$. $g(x)$ will typically be $\log(x)$ if x is defined to be positive, or $g(x) = x$ if x is unrestricted. PROPORTIONAL HAZARDS / ACCELERATED FAILURE TIME

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 (or delayed-entry) times, meaning that individual i is only observed conditionally on having survived up to s_i, thus $s_i = 0$ if there is no left-truncation. Additionally let t_i^{max} be left-censoring times. If there is no left-censoring then these are infinite, so that $S(t_i^{max}) = 0$; or if the ith death time is interval-censored then $c_i = 0$ and t_i^{max} is finite.

The likelihood for the full set of parameters $\boldsymbol{\gamma}$ in model (1), given the corresponding data vectors, is

$$l(\boldsymbol{\gamma}|\mathbf{t}, \mathbf{c}, \mathbf{s}, \mathbf{t}^{max}) = \left\{ \prod_{i:\ c_i=1} f_i(t_i) \prod_{i:\ c_i=0} (S_i(t_i) - S_i(t_i^{max})) \right\} / \prod_i S_i(s_i) \tag{2}$$

Note that the individuals are independent, so that **flexsurv** does not currently support frailty, clustered or random effects models.

EXAMPLE DATASET HERE. bc? ovarian, lung, cancer, aml, tobin, cgd, heart, kidney, logan, nwtco, pbc, rats in survival

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 gives the response as a survival object using `Surv` function from the **survival** package. Here, this indicates that the response variable is `recyrs`, and that these are observed death and censoring times when the variable `censrec` is 1 or 0 respectively. All of these variables are in the data frame called `bc`.

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

If we also had left-truncation times in a variable called `start`, the response would be `Surv(start,recyrs,censrec)`. Or if all responses were interval-censored between lower and upper bounds `tmin` and `tmax`, then we would write `Surv(tmin,tmax,type="interval2")`.

3.1. Using a built-in survival model

If the argument `dist` is a string, this denotes a built-in survival distribution. The currently built-in distributions are listed in Table 1. In each case, the probability density $f()$ and parameters used in the fitted model is taken from an existing R function of the same name but beginning with the letter `d`. For example if `dist="weibull"`, the density function is `dweibull`.

For the Weibull, exponential (`dexp`), gamma (`dgamma`) and log-normal (`dlnorm`), the density functions are provided with standard installations of R. **flexsurv** provides some additional survival distributions including the Gompertz distribution with unrestricted shape parameter (`dist="gompertz"`), and two more flexible families:

Generalised gamma This three-parameter distribution includes the Weibull, gamma and log-normal as special cases. The parameterisation from Stacy (1962) is available as `dist="gengamma.orig"`, however the newer parameterisation Prentice (1974) is preferred for modelling since the log-normal is not at the boundary of the parameter space, and includes a further class of mdoels with negative Q.

See `help(GenGamma)` and `help(GenGamma.orig` for the exact definitions of these distributions and which parameter values correspond to the Weibull, gamma or log-normal. WRITE OUT IN STATA DEFINITION?

Generalized F This four-parameter distribution includes the generalized gamma, and also the log-logistic, as special cases. The variety of hazard shapes that can be represented is discussed by Cox (2008). It is provided here in alternative "original" (`dist="genf.orig"`) and "stable" parameterisations (`dist="genf"`) as discussed by ?. Again, see `help(GenF)` and `help(GenF.orig` for the exact definitions.

	Parameters	Density R function	`dist`
Exponential	rate	dexp	`exp`
Weibull	shape, scale	dweibull	`weibull`
Gamma	shape, rate	dgamma	`gamma`
Log-normal	meanlog, sdlog	dlnorm	`lnorm`
Gompertz	shape, rate	dgompertz	`gompertz`
Generalized gamma (Prentice 1975)		dgengamma	`gengamma`
Generalized gamma (Stacy 1962)		dgengamma	`gengamma.orig`
Generalized F (stable)		dgenf	`genf`
Generalized F (original)		dgenf	`genf.orig`

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

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

3.2. Supplying own distributions

flexsurv is not limited to its built-in distributions. Any survival model of the form (1–2) can be fitted if we can provide either the density function $f()$ or the hazard $h()$. Many contributed R packages provide probability density and cumulative distribution functions for positive distributions. On the other hand, survival models are naturally specified by through their hazard function, representing the changing risk of death through time. For example, for survival following major surgery we may want a "U-shaped" hazard curve, representing a high risk soon after the operation, which then decreases, but increases naturally as survivors grow older.

Example: Using functions from a contributed package Distribution exists in another package, but may be parameterised Example: Gompertz-Makeham Functions need to be vectorised

Example: Changing the parameterisation of a distribution (talk about Weibull prop haz model, GG PH) refer back to GG presentation

Example: Omitting the cumulative distribution or hazard If there is no analytic form for $F(t)$ or $H(t)$ as the integral of the density or hazard respectively, then **flexsurv** can compute these internally by numerical integration, though this will substantially slow down the computation. The default options of the built-in R routine `integrate` for adaptive quadrature are used, though these may be changed using the `integ.opts` argument to `flexsurvreg`.

In Section 4.1

3.3. Computation

The likelihood is maximised using the optimisation methods available through the standard R `optim` function. By default, this is the `"BFGS"` method ((Nash 1990)) which can use the analytic derivatives of the likelihood with respect to the model parameters, if these are available, to improve the speed of convergence to the maximum.

For custom distributions, the user can optionally supply functions with names beginning "DLd" and "DLS" respectively (e.g. DLdweibull,DLSweibull) to calculate the derivatives of the log density and log survivor functions with respect to the transformed parameters γ.

Initial values are difficult: ideally two would come from moments of the distribution, then defaults that reduce to simpler distributions. example

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

3.4. 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

4. 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

4.1. 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

5. Multi-state models

enhances mstate

6. Potential extensions

relative survival frailty what else does survival do many extensions may come from user-contributed models

A. Acknowledgements

Thanks to Milan Bouchet-Valat.

References

BrostrÃűm G (2014). *eha: Event History Analysis.* R package version 2.4-1, URL http://CRAN.R-project.org/package=eha.

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

Crowther MJ, Lambert PC (2013). "stgenreg: A Stata package for general parametric survival analysis." *Journal of Statistical Software*, **53**, 1–17.

Lambert PC, Royston P (2009). "Further development of flexible parametric models for survival analysis." *Stata Journal*, **9**(2), 265.

Latimer NR (2013). "Survival analysis for economic evaluations alongside clinical trialsâĂŤextrapolation with patient-level data inconsistencies, limitations, and a practical guide." *Medical Decision Making*, **33**(6), 743–754.

Nadarajah S, Bakar S (2013). "A new R package for actuarial survival models." *Computational Statistics*, **28**(5), 2139–2160.

Nash JC (1990). *Compact numerical methods for computers: linear algebra and function minimisation.* CRC Press.

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

Reid N (1994). "A conversation with Sir David Cox." *Statistical Science*, pp. 439–455.

Royston P, Parmar M (2002). "Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects." **21**(1), 2175–2197.

Stacy EW (1962). "A generalization of the gamma distribution." *Annals of Mathematical Statistics*, (33), 1187–92.

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

Yee TW, Wild C (1996). "Vector generalized additive models." *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 481–493.