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

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

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 $h(t) = f(t)/S(t)$.

$\mu = \alpha_0$ 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.

Covariates All parameters may depend on a vector of covariates $\mathbf{z}$ through link-transformed linear models $g_0(\mu) = \boldsymbol{\beta}_0'\mathbf{z}$ and $g_r(\alpha_r) = \boldsymbol{\beta}_r'\mathbf{z}$. $g(x)$ will typically be $\log(x)$ if x is defined to be positive, or $g(x) = x$ if x is unrestricted. In all models, $\boldsymbol{\beta}$ includes at least an intercept, so that the full set of parameters is given by $\{\boldsymbol{\beta}_r : r = 1, \ldots, R$.

Suppose that the location but not the ancillary parameters depend on covariates. If the hazard function factorises as $h(t|\alpha, \mu(\mathbf{z})) = \mu(\mathbf{z})h_0(t|\alpha)$, then this is a *proportional hazards* model, so that the hazard ratio between two groups (defined by different values of $\mathbf{z}$ is constant over time. Alternatively, if $S(t|\mu(\mathbf{z}), \alpha) = S(\mu(\mathbf{z})t|\alpha)$ then we have an *accelerated failure time* model, so that the effect of covariates is to speed or slow the passage of time. For example if there is one covariate with coefficient $\beta = \log(2)$, then doubling the covariate value would give half the expected survival time.

Data and likelihood 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 parameters $\boldsymbol{\beta}$ in model (1), given the corresponding data vectors, is

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

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

Illustrate survreg and flexsurvreg, par ests come from dweibull.

flexsurv provides some additional survival distributions including the Gompertz distribution with unrestricted shape parameter (`dist="gompertz"`), and two more flexible families:

Generalized gamma This three-parameter distribution includes the Weibull, gamma and log-normal as special cases. The original parameterisation from Stacy (1962) is available as

	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**

`dist="gengamma.orig"`, however the newer parameterisation (Prentice 1974) is preferred: `dist="gengamma"`. This has parameters (μ,σ,q), and survivor function

$$\begin{array}{ll} 1 - I(\gamma, u) & (q > 0) \\ 1 - \Phi(z) & (q = 0) \end{array}$$

where $I(a, x) = \int_0^x x^{a-1} \exp(-x)/\Gamma(a)$ is the incomplete gamma function (the cumulative gamma distribution with shape a and scale 1), Φ is the standard normal cumulative distribution, $u = \gamma \exp(|q|z)$, $z = (\log(t) - \mu)/\sigma$, and $\gamma = q^{-2}$. The Prentice (1974) parameterisation extends the original one to include a further class of models with negative q, and survivor function $I(\gamma, u)$, where z is replaced by $-z$. This stabilises estimation when the distribution is close to log-normal, since $q = 0$ is no longer near the boundary of the parameter space. In R notation, [1] the parameter values corresponding to the three special cases are

```
dgengamma(x, mu, sigma, Q=0)     ==  dlnorm(x, mu, sigma)
dgengamma(x, mu, sigma, Q=1)     ==  dweibull(x, shape=1/sigma, scale=exp(mu))
dgengamma(x, mu, sigma, Q=sigma) ==  dgamma(x, shape=1/sigma^2,
                                            rate=exp(-mu) / sigma^2)
```

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 presented by `?`. See `help(GenF)` and `help(GenF.orig)` in the package documentation for the exact definitions.

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,

[1] The parameter called q here and in previous literature is called Q in `dgengamma` and related functions, since the first argument of a cumulative distribution function is conventionally named `q`, for quantile, in R.

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 Weibull model presentation 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 4.2

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. Any-dimension models

flexsurv also supports models where the number of parameters is arbitrary. In the models discussed previously, the number of parameters in the model family is fixed (e.g. three for the generalized gamma). Here the model complexity can be chosen by the user. We may want to represent more irregular hazard curves by more flexible functions, or use bigger models if a bigger sample size makes it feasible to estimate more parameters.

4.1. Royston and Parmar spline model

In the spline-based survival model of Royston and Parmar (2002), a transformation $g(S(t,z))$ of the survival function is modelled as a natural cubic spline function of log time, $x = \log(t)$, plus linear effects of covariates z. This is available here as the function `flexsurvspline`, and is also available in the Stata package `stpm2` (Lambert and Royston 2009) (and historically `stpm`, Royston (2001, 2004)).

$$g(S(t,z)) = s(x, \boldsymbol{\gamma})$$

Typically we use $g(S(t,\mathbf{z}) = \log(-\log(S(t,\mathbf{z}))) = \log(H(t,\mathbf{z}))$, the log cumulative hazard, giving a proportional hazards model.

Spline parameterisation The complexity of the model, thus the dimension of $\boldsymbol{\gamma}$, is governed by the number of *knots* m in the spline function $s()$. Natural cubic splines are piecewise cubic polynomials defined to be continuous, with continuous first and second derivatives at the knots, and also constrained to be linear beyond boundary knots k_{min}, k_{max}. As well as the boundary knots there may be up to $m \geq 0$ *internal* knots $k_1, \ldots k_m$. Various spline parameterisations exist — the one used here is from Royston and Parmar (2002).

$$s(x, \boldsymbol{\gamma}) = \gamma_0 + \gamma_1 x + \gamma_2 v_1(x) + \ldots + \gamma_{m+1} v_m(x) \tag{3}$$

where $v_j(x)$ is the jth *basis* function

$$v_j(x) = (x - k_j)_+^3 - \lambda_j (x - k_{min})_+^3 - (1 - \lambda_j)(x - k_{max})_+^3, \qquad \lambda_j = \frac{k_{max} - k_j}{k_{max} - k_{min}}$$

and $(x-a)_+ = max(0, x-a)$. If $m = 0$ then there are only two parameters γ_0, γ_1 — in fact if $g()$ is the log cumulative hazard, this is equivalent to a Weibull model (PARAMETERS / DPQR STATEMENT). Table 2 explains two alternative choices of $g()$.

Covariates on spline parameters Covariates can be placed on any parameter γ through a linear model (with identity link function). Most straightforwardly we can let the intercept γ_0 vary with covariates $\mathbf{z}$, giving a proportional hazards or odds model (depending on $g()$).

$$g(S(t,z)) = s(x, \boldsymbol{\gamma}) + \boldsymbol{\beta}^T \mathbf{z}$$

Model	$g(S(t,\mathbf{z}))$	In `flexsurvspline`	With $m=0$
Proportional hazards	$\log(-\log(S(t,\mathbf{z})))$ (log cumulative hazard)	`scale="hazard"`	Weibull
Proportional odds	$\log(S(t,\mathbf{z})^{-1}-1)$ (log cumulative odds)	`scale="odds"`	Log-logistic
Normal / probit	$\Phi^{-1}(S(t,\mathbf{z}))$ (inverse normal CDF, `qnorm`)	`scale="normal"`	Log-normal

Table 2: Alternative modelling scales for `flexsurvspline`

FLEXSURVSPLINE COMMAND

The spline coefficients $\gamma_j : j = 1, 2 \ldots$, the "ancillary parameters", may also be modelled as linear functions of covariates $\mathbf{z}$, as

$$\gamma_j(\mathbf{z}) = \gamma_{j0} + \gamma_{j1} z_1 + \gamma_{j2} z_2 + \ldots$$

giving a model where the effects of covariates are arbitrarily flexible functions of time: a non-proportional hazards or odds model.

FLEXSURVSPLINE COMMAND

Demo on at least one dataset

DPQR STATEMENTS

4.2. General-dimension models

The spline model above is an example of the general parametric form (1), but the number of parameters ($R+1$ in (1), $m+2$ in (3)) is arbitrary. **flexsurv** has the tools to deal with any model of this form. `flexsurvspline` works internally by building a custom distribution and then calling `flexsurvreg`. Similar models may in principle be built by users using the same method. This relies on a functional programming trick.

Creating distribution functions dynamically The R distribution functions supplied to custom models are expected to have a fixed number of arguments, one for each scalar parameter. However, the distribution functions for the spline model (e.g. dsurvspline) have an argument `gamma` representing the vector of parameters γ.

To convert it into the correct form, **flexsurv** provides the utility `unroll.function`. This converts a function with one (or more) vector parameters (matrix arguments) to a function with an arbitrary number of scalar parameters (vector arguments).

Note the scalar *parameters* can be supplied to the function as correspond to vector *arguments*, and vector parameters supplied as *matrix* arguments.

That is, `x`,`shape` and `scale` in `dweibull(x,shape,scale)` could actually be vectors representing alternative values for the `x` or the parameters. Similary `gamma` in `dsurvspline(x, gamma,...)` could be a matrix, whose rows represent alternative sets of values for γ.

Due to vectorisation: EXAMPLE

Example: splines on alternative scales 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).

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

5. Multi-state models

A *multi-state model* represents how an individual moves between multiple states through time. Survival analysis is a special case of multi-state modelling with two states "alive" and "dead". Suppose an individual is in state $S(t)$ at time t. The next state to which the individual moves, and the time of the change, are governed by a set of *transition intensities* $q_{rs}(t)$ for states $r, s = 1, \ldots, R$, which for a survival model are equivalent to the hazard $h(t)$. The intensity represents the instantaneous risk of moving from state r to state s.

Suppose our data consist of a series of event times $t_1, \ldots, t_n$, the last of these may be an observed event or censoring. Any software to fit survival models can also fit multi-state models to this kind of data, provided it can deal with left-truncation or *counting process* data.

For more discussion of the theory see Putter, Fiocco, and Geskus (2007). ref also Andersen for CP data

Counting process data For each permitted $r \to s$ transition in the multi-state model (ILLUSTRATION) there is a corresponding *time-to-event model*, with cause-specific hazard rates defined by $q_{rs}(t)$. To enable estimation of these hazards, the data are expressed as a series of times to events which are potentially censored: $dt_j = t_{j+1} - t_j : j = 1, \ldots, n-1$. For a patient who moves into state s at time t_j, their next event at t_{j+1} is defined by the model structure (Figure~**??**) to be one of a set of competing events $s^*_1, \ldots, s^*_{n_s}$.

For example, in state EXAMPLE, the next state must either be EXAMPLE or EXAMPLE so $n_s = EG$. The time of the event which actually occurs at t_{j+1} is *observed*, and the times of the *competing* events from this set (which have not occurred by this time) are *censored*. Each dt_j contributes an *observed* time to one of the EG transition-specific models, and a *censored* time to each of the models for the competing events.

FLEXSURVREG EXAMPLE

The **mstate** R package (de~Wreede, Fiocco, and Putter 2010, 2011) has a utility `msprep` to produce data of this form from "wide-format" datasets where rows represent individuals, and times of different events appear in different columns, and **mstate** has a utility `msm2Surv` for

illustrates Cox models flexible parametric multi-state models

Prediction from multi-state models Define cumulative incidence functions

The `mstate` package is designed to work with piecewise-constant cumulative incidence functions baseline hazards are estimated non-parametrically (de~Wreede *et~al.* 2010, 2011)

function `msfit` that produces the cumulative incidences for each transition and a given covariate category, and their covariances, given a Cox model fitted using `coxph` from the **survival** package.

Aalen-Johansen estimator, simulation

contrast Markov and semi-Markov models

Multi-state models for panel data Note the contrast with multi-state models for *panel data*, that is, observations of the state of the process at a series of times (Kalbfleisch and Lawless 1985). In panel data, we do not necessarily know know the time of each transition, or even whether a transitions of a certain type have occurred at all between a pair of observations. Such models can be fitted with the **msm** package for R, but are restricted to (piecewise) exponentially-distributed event times.

`survSplit` function in **survival**

6. Potential extensions

relative survival frailty 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.

de˜Wreede L, Fiocco M, Putter H (2010). "The `mstate` package for estimation and prediction in non-and semi-parametric multi-state and competing risks models." *Computer Methods and Programs in Biomedicine*, **99**(3), 261–274.

de˜Wreede LC, Fiocco M, Putter H (2011). "`mstate`: an R package for the analysis of competing risks and multi-state models." *J Stat Softw*, **38**, 1–30.

Kalbfleisch J, Lawless J (1985). "The Analysis of Panel Data under a Markov Assumption." *Journal of the American Statistical Association*, **80**(392), 863–871.

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.

Putter H, Fiocco M, Geskus RB (2007). "Tutorial in biostatistics: competing risks and multi-state models." **26**, 2389–2430.

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

Royston P (2001). "Flexible parametric alternatives to the Cox model, and more." *Stata Journal*, **1**(1), 1–28.

Royston P (2004). "Flexible parametric alternatives to the Cox model: update." *The Stata Journal*, **4**(1), 98–101.

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.

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