

# longevity: An R Package for Modelling Excess Lifetimes

by Léo R. Belzile

**Abstract** The longevity R package provides a maximum likelihood estimation routine for modelling of survival data that are subject to non-informative censoring or truncation. It includes a selection of 12 parametric models of varying complexity, with a focus on tools for extreme value analysis and more specifically univariate peaks over threshold modelling. The package provides utilities for univariate threshold selection, parametric and nonparametric maximum likelihood estimation, goodness of fit diagnostics and model comparison tools. These different methods are illustrated using individual Dutch records and aggregated Japanese human lifetime data.

## 1 Introduction

Many data sets collected by demographers for the analysis of human longevity have unusual features and for which limited software implementations exist. The **longevity** package was initially built for dealing with human longevity records and data from the International Database on Longevity (IDL), which provides age at death of supercentenarians, i.e., people who died above age 110. Data for the statistical analysis of (human) longevity can take the form of aggregated counts per age at death, or most commonly life trajectory of individuals with both birth and death dates. Such lifetimes are often interval truncated (only age at death of individuals dying between two calendar dates are recorded) or left truncated and right censored (when data of individuals still alive at the end of the collection period are also included). Another frequent format is death counts, aggregated per age band. Censoring and truncation are typically of administrative nature and thus non-informative about death.

Supercentenarians are extremely rare and records are sparse. The most popular parametric models used by practitioners are justified by asymptotic arguments and have their roots in extreme value theory. Univariate extreme value distributions are well implemented in software and Belzile et al. (2023) provides a recent review of existing implementations. While there are many standard R packages for the analysis of univariate extremes using likelihood-based inference, such as **evd** (Stephenson, 2002), **mev** and **extRemes** (Gilleland and Katz, 2016), only the **evgam** package includes functionalities to fit threshold exceedance models with censoring, as showcased in Youngman (2022) with rounded rainfall measurements. Support for survival data for extreme value models is in general wholly lacking, which motivated the development of **longevity**.

The **longevity** package also includes several parametric models commonly used in demography. Many existing packages that focus on tools for modelling mortality rates, typically through life tables, are listed in the CRAN Task View *ActuarialScience* in the Life insurance section. They do not however allow for truncation or more general survival mechanisms, as the aggregated data used are typically complete except for potential right censoring for the oldest age group. The **survival** package (Therneau and Grambsch, 2000) includes utilities for accelerated failure time models for 10 parametric models. The **MortCast** package can be used to estimate age-specific mortality rates using the Kannisto and Lee-Carter approaches, among others (Ševčíková et al., 2016). The **demography** package provides forecasting methods for death rates based on constructed life tables using the Lee-Carter or ARIMA models. **MortalityLaws** includes utilities to download data from the Human Mortality Database (HMD), and fit a total of 27 parametric models for life table data (death counts and population at risk per age group) using Poisson, binomial or alternative loss functions. The **vitality** package fits the family of vitality models (Li and Anderson, 2009; Anderson, 2000) via maximum likelihood based on empirical survival data. The **fitdistrplus** (Delignette-Muller and Dutang, 2015) allows for generic parametric distributions to be fitted to interval censored data via maximum likelihood, with various S3 methods for

model assessment. The package also allows user-specified models, thereby permitting custom definitions for truncated distributions whose truncation bounds are passed as fixed vector parameters. Parameter uncertainty can be obtained via additional functions using nonparametric bootstrap. Many parametric distributions also appear in the **VGAM** package (Yee and Wild, 1996; Yee, 2015), which allows for vector generalized linear modelling. The **longevity** package is less general and offers support only for selected parametric distributions, but contrary to the aforementioned packages allows for truncation and general patterns. One strength of **longevity** is that it also includes model comparison tools that account for non-regular asymptotics and goodness of fit diagnostics.

Nonparametric methods are popular tools for the analysis of survival data with large samples, owing to their limited set of assumptions. They also serve for the validation of parametric models. Without explanatory variable, a closed-form estimator of the nonparametric maximum likelihood estimator of the survival function can be derived in particular instances, including the product limit estimator (Kaplan and Meier, 1958) for the case of random or non-informative right censoring and an extension allowing for left truncation (Tsai et al., 1987). In general, the nonparametric maximum likelihood estimator of the survival function needs to be computed using an expectation-maximization (EM) algorithm (Turnbull, 1976). Nonparametric estimators only assign probability mass on observed failure times and intervals and so cannot be used for extrapolation beyond the range of the data, limiting its utility in extreme value analysis.

The CRAN Task View on *Survival* Analysis lists various implementations of nonparametric maximum likelihood estimators of the survival or hazard functions: **survival** implements the Kaplan–Meier and Nelson–Aalen estimators. Many packages focus on the case of interval censoring (Groeneboom and Wellner, 1992, §3.2), including **prodlm**; Anderson-Bergman (2017a) reviews the performance of the implementations in **icenReg** (Anderson-Bergman, 2017b) and **lcens**. The latter uses the incidence matrix as input data. Routines for doubly censored data are provided by **dblcens**. The **interval** (Fay and Shaw, 2010) implements Turnbull’s EM algorithm for interval censored data. The case of left truncated and right censored data is handled by **survival**, and **tranSurv** provides transformation models with the potential to account for truncation dependent on survival times. For interval truncated data, dedicated algorithms that use gradient-based steps (Efron and Petrosian, 1999) or inverse probability weighting (Shen, 2010) exist and can be more efficient than the EM algorithm of Turnbull (1976). Many of these are implemented in the **DTDA** package. The **longevity** package includes a C++ implementation of the corrected Turnbull’s algorithm (Turnbull, 1976), returning the nonparametric maximum likelihood estimator for arbitrary censoring and truncation patterns, as opposed to the specific existing aforementioned implementations already available which focus on specific subcases. With a small number of observations, it is also relatively straightforward to maximize the log likelihood for the concave program subject to linear constraints using constrained optimization algorithms; **longevity** relies for this on **Rsolnp**, which uses augmented Lagrangian methods and sequential quadratic programming (Ye, 1987).

### 1.1 Motivating examples

To showcase the functionality of the package and particularity for the modelling of threshold exceedances, we consider Dutch and Japanese life lengths. The dutch database contains the age at death (in days) of Dutch people who died above age 92 between 1986 and 2015; these data were obtained from Statistics Netherlands and analyzed in Einmahl et al. (2019) and Belzile et al. (2022). Records are interval truncated, as people are included in the database only if they died during the collection period. In addition, there are 226 interval censored and interval truncated records for which only the month and year of birth and death are known, as opposed to exact dates.

The second database we consider is drawn from Maier et al. (2021). The data frame `japanese2` consists of counts of Japanese above age 100 by age band and are stratified by both birth cohort and sex. To illustrate the format of the data, counts for female Japanese

**Table 1:** Death count by birth cohort and age band for female Japanese.

age	1874-1878	1879-1883	1884-1888	1889-1893	1894-1898	1899-1900
100	1648	2513	4413	8079	16036	9858
101	975	1596	2921	5376	11047	7091
102	597	987	1864	3446	7487	4869
103	345	597	1153	2230	5014	3293
104	191	351	662	1403	3242	2133
105	121	197	381	855	2084	1357
106	64	122	210	495	1284	836
107	34	74	120	274	774	521
108	16	41	66	152	433	297
109	12	30	39	83	252	167
110	6	17	21	49	130	92
111	4	10	15	26	69	47
112	3	3	11	15	29	22
113	2	2	8	5	15	9
114	1	2	4	3	7	2
115	0	1	1	0	3	2
116	0	1	1	0	1	1
117	0	0	0	0	1	1

are reproduced in Table 1. The data were constructed using the extinct cohort method and are interval censored between age and age + 1 and right truncated at the age reached by the oldest individuals of their birth cohort in 2020. The count variable lists the number of instances in the contingency table, and serves as a weight for likelihood contributions.

## 2 Package functionalities

The **longevity** package uses the S3 object oriented system and provides a series of functions with common arguments. The syntax used by the **longevity** package purposely mimics that of the **survival** package (Therneau and Grambsch, 2000), except that it does not specify models using a formula. Users must provide vectors for the time or age (or bounds in case of intervals) via arguments `time` and `time2`, as well as lower and upper truncation bounds (`l trunc` and `r trunc`) if applicable. The integer vector `event` is used to indicate the type of event, where following **survival** 0 indicates right-censoring, 1 observed events, 2 left-censoring and 3 interval censoring. Together, these five vectors characterize the data and the survival mechanisms at play. Depending on the sampling scheme, not all arguments are required or relevant and they need not be of the same length, but are common to most functions. Users can also specify a named list `args` to pass arguments: as illustrated below, this is convenient to avoid specifying repeatedly the common arguments in each function call. Default values are overridden by elements in `args`, with the exception of those that are passed by the user directly in the call. Relative to **survival**, functions have additional arguments `l trunc` and `r trunc` for left and right truncation limits, as these are also possibly matrices for the case of double interval truncation (Belzile et al., 2022), since both censoring and truncation can be present simultaneously.

We can manipulate the data set to build the time vectors and truncation bounds for the Dutch data. We re-scale observations to years for interpretability and keep only records above age 98 for simplicity. We split the data to handle the observed age at death first: these are treated as observed (uncensored) whenever `time` and `time2` coincide. When exact dates are not available, we compute the range of possible age at which individuals may have died, given their birth and death years and months. The truncation bounds for each individual can be obtained by subtracting from the endpoints of the sampling frame the birth dates,

**Table 2:** Sample of five Dutch records, formatted so that the inputs match the function arguments used by the package. Columns give the age in years at death (or plausible interval), lower and upper truncation bounds giving minimum and maximum age for inclusion, an integer indicating the type of censoring, gender and birth year.

time	time2	ltrunc	rtrunc	event	gender	byear
104.67	105.74	80.00	111.00	3	female	1905
103.50	104.58	78.00	109.00	3	female	1907
100.28	100.28	92.01	104.50	1	female	1911
100.46	100.46	92.01	102.00	1	male	1913
100.51	100.51	92.01	104.35	1	female	1911

with left and right truncation bounds

$$\text{ltrunc} = \min\{92 \text{ years}, 1986.01.01 - \text{bdate}\}, \quad \text{rtrunc} = 2015.12.31 - \text{bdate}.$$

Table 2 shows a sample of five individuals, two of whom are interval-censored, and the corresponding vectors of arguments along with two covariates, gender (gender) and birth year (byear).

We can proceed similarly for the Japanese data. Ages of centenarians are rounded down to the nearest year, so all observations are interval censored within one-year intervals. Assuming that the ages at death are independent and identically distributed with distribution function  $F(\cdot; \theta)$ , the log likelihood for exceedances  $y_i = \text{age}_i - u$  above age  $u$  is

$$\ell(\theta) = \sum_{i: \text{age}_i > u} n_i [\log\{F(y_i + 1; \theta) - F(y_i; \theta)\} - \log F(r_i - u; \theta)]$$

where  $n_i$  is the count of the number of individuals in cell  $i$  and  $r_i > \text{age}_i + 1$  is the right truncation limit for that cell, i.e., the maximum age that could have been achieved for that birth cohort by the end of the data collection period.

```
data(japanese2, package = "longevity")
# Keep only non-empty cells
japanese2 <- japanese2[japanese2$count > 0, ]
# Define arguments that are recycled
japanese2$rtrunc <- 2020 -
  as.integer(substr(japanese2$bcohort, 1, 4))
# The line above extracts the earliest year of the birth cohort
# Create a list with all arguments common to package functions
args_japan <- with(japanese2,
  list(
    time = age, # lower censoring bound
    time2 = age + 1L, # upper censoring bound
    event = 3, # define interval censoring
    type = "interval2",
    rtrunc = rtrunc, # right truncation limit
    weights = count)) # counts as weights
```

## 2.1 Parametric models and maximum likelihood estimation

Various models are implemented in [longevity](#): their hazard functions are reported in Table 3. Two of those models, labelled perks and beard are logistic-type hazard functions proposed in [Perks \(1932\)](#) that have been used by [Beard \(1963\)](#), and popularized in work of Kannisto and Thatcher; we use the parametrization of [Richards \(2012\)](#), from which we also adopt the nomenclature. Users can compare the models with those available in [MortalityLaws](#); see `?availableLaws` for the list of hazard functions and parametrizations.

**Table 3:** List of parametric models for excess lifetime supported by the package, with parametrization and hazard functions.

model	hazard function	constraints
exp	$\sigma^{-1}$	$\sigma > 0$
gomp	$\sigma^{-1} \exp(\beta t / \sigma)$	$\sigma > 0, \beta \geq 0$
gp	$(\sigma + \xi t)^{-1}$	$\sigma > 0, \xi \in \mathbb{R}$
weibull	$\sigma^{-\alpha} \alpha t^{\alpha-1}$	$\sigma > 0, \alpha > 0$
extgp	$\beta \sigma^{-1} \exp(\beta t / \sigma) [\beta + \xi \{\exp(\beta t / \sigma) - 1\}]^{-1}$	$\sigma > 0, \beta \geq 0, \xi \in \mathbb{R}$
extweibull	$\alpha \sigma^{-\alpha} t^{\alpha-1} \{1 + \xi (t / \sigma)^\alpha\}_+$	$\sigma > 0, \alpha > 0, \xi \in \mathbb{R}$
perks	$\alpha \exp(\nu x) / \{1 + \alpha \exp(\nu x)\}$	$\nu \geq 0, \alpha > 0$
beard	$\alpha \exp(\nu x) / \{1 + \alpha \beta \exp(\nu x)\}$	$\nu \geq 0, \alpha > 0, \beta \geq 0$
gompmake	$\lambda + \sigma^{-1} \exp(\beta t / \sigma)$	$\lambda \geq 0, \sigma > 0, \beta \geq 0$
perksmake	$\lambda + \alpha \exp(\nu x) / \{1 + \alpha \exp(\nu x)\}$	$\lambda \geq 0, \nu \geq 0, \alpha > 0$
beardmake	$\lambda + \alpha \exp(\nu x) / \{1 + \alpha \beta \exp(\nu x)\}$	$\lambda \geq 0, \nu \geq 0, \alpha > 0, \beta \geq 0$

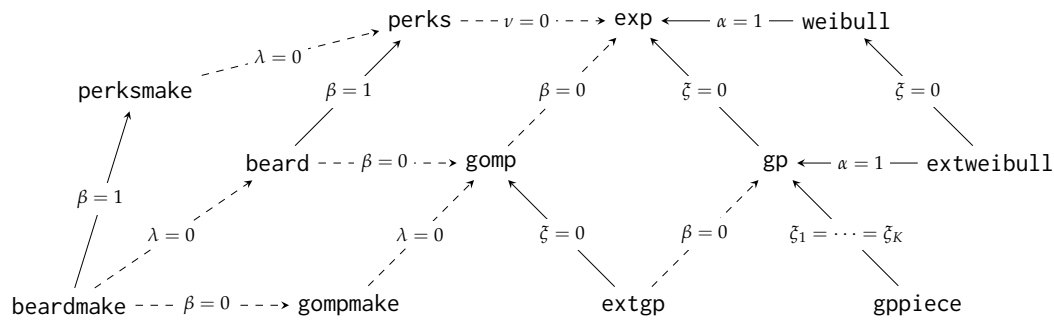
Many of the models are nested and Figure 1 shows the logical relation between the various families. The function `fit_elife` allows users to fit all of the parametric models of Table 3: the `print` method returns a summary of the sampling mechanism, the number of observations, the maximum log likelihood and parameter estimates with standard errors. Depending on the data, some models may be overparametrized and parameters need not be numerically identifiable. To palliate such issues, the optimization routine, which uses **Rsolnp**, can try multiple starting values or fit various sub-models to ensure that the parameter values returned are indeed the maximum likelihood estimates. If one tries to compare nested models and the fit of the simpler model is better than the alternative, the `anova` function will return an error message.

The `fit_elife` function handles arbitrary censoring patterns over single intervals, along with single interval truncation and interval censoring. To accommodate the sampling scheme of the International Database on Longevity (IDL), an option also allows for double interval truncation (Belzile et al., 2022), whereby observations are included only if the person dies between time intervals, potentially overlapping, which defines the observation window over which dead individuals are recorded.

```

thresh <- 108
model0 <- fit_elife(arguments = args_japan,
  thresh = thresh,
  family = "exp")
(model1 <- fit_elife(arguments = args_japan,
  thresh = thresh,
  family = "gomp"))
#> Model: Gompertz distribution.
#> Sampling: interval censored, right truncated
#> Log-likelihood: -3599.037
#>
#> Threshold: 108
#> Number of exceedances: 2489
#>
#> Estimates
#> scale shape
#> 1.6855 0.0991
#>
#> Standard Errors
#> scale shape
#> 0.0523 0.0273
#>

```



**Figure 1:** Relationship between parametric models showing nested relations. Dashed arrows represent restrictions that lead to nonregular asymptotic null distribution for comparison of nested models. Comparisons between models with Makeham components and exponential are not permitted by the software because of nonidentifiability issues.

```
#> Optimization Information
#> Convergence: TRUE
```

## 2.2 Model comparisons

Goodness of fit of nested models can be compared using likelihood ratio tests via the anova method. Most of the interrelations between models yield non-regular model comparisons since, to recover the simpler model, one must often fix parameters to values that lie on the boundary of the parameter space. For example, if we compare a Gompertz model with the exponential, the limiting null distribution is a mixture of a point mass at zero and a  $\chi^2_1$  variable, both with probability half (Chernoff, 1954). Many authors (e.g., Camarda, 2022) fail to recognize this fact. The case becomes more complicated with more than one boundary constraint: for example, the deviance statistic comparing the Beard–Makeham and the Gompertz model, which constrains two parameters on the boundary of the parameter space, has a null distribution which is a mixture of  $\chi^2_2/4 + \chi^2_1/2 + \chi^2_0/4$  (Self and Liang, 1987).

Nonidentifiability impacts testing: for example, if the rate parameter of the Perks–Makeham model (perksmake)  $\nu \rightarrow 0$ , the limiting hazard,  $\lambda + \exp(\alpha) / \{1 + \exp(\alpha)\}$ , is constant (exponential model), but neither  $\alpha$  nor  $\lambda$  is identifiable. The usual asymptotics for the likelihood ratio test break down as the information matrix is singular (Rotnitzky et al., 2000). As such, all three families that include a Makeham component cannot be directly compared to the exponential in **longevity** and the call to anova returns an error message.

Users can also access information criteria, AIC and BIC. The correction factors implemented depend on the number of parameters of the distribution, but do not account for singular fit, non-identifiable parameters or singular models for which the usual corrections  $2p$  and  $\ln(n)p$  are inadequate (Watanabe, 2010).

To showcase how hypothesis testing is performed, we consider a simple example with two nested models. We test whether the exponential model is an adequate simplification of the Gompertz model for exceedances above 108 years — an irregular testing problem since  $\beta = 0$  is a restriction on the boundary of the parameter space. The drop in log likelihood is quite large, indicating the exponential model is not an adequate simplification of the Gompertz fit. This is also what is suggested by the Bayesian information criterion, which is much lower for the Gompertz model than for the exponential.

```
# Model comparison
anova(model1, model0)
# Information criteria
c("exponential" = BIC(model0), "Gompertz" = BIC(model1))
```



	npar	Deviance	Df	Chisq	Pr(>Chisq)
gomp	2	7198.07			
exp	1	7213.25	1	15.17	0

```
#> exponential      Gompertz
#>      7221.065      7213.713
```

### 2.3 Simulation-based inference

Given the poor finite sample properties of the aforementioned tests, it may be preferable to rely on a parametric bootstrap rather than on the asymptotic distribution of the test statistic (Belzile et al., 2022) for model comparison. Simulation-based inference requires capabilities for drawing new data sets whose features match those of the original one. For example, the [International Database on Longevity](#) (IDL) (Jdanov et al., 2021) features data that are interval truncated above 110 years, but doubly interval truncated since the sampling period for semisupercentenarians (who died age 105 to 110) and supercentenarians (who died above 110) are not always the same (Belzile et al., 2022). The 2018 Istat semisupercentenarians database analyzed by Barbi et al. (2018) on Italians includes left truncated right censored records.

To mimic the postulated data generating mechanism while accounting for the sampling scheme, we could use the observed birth dates, or simulate new birth dates (possibly through a kernel estimator of the empirical distribution of birth dates) while keeping the sampling frame with the first and last date of data collection to define the truncation interval. In other settings, one could obtain the nonparametric maximum likelihood estimator of the distribution of the upper truncation bound (Shen, 2010) using an inverse probability weighted estimator, which for fixed data collection windows is equivalent to setting the birth date.

The `samp_elif` function includes multiple `type2` arguments to handle these. For interval truncated data (`type2="ltrt"`), it uses the inversion method (Section 2 of Devroye (1986)): for  $F$  an absolutely continuous distribution function and  $F^{-1}$  the corresponding quantile function, a random variable distributed according to  $F$  truncated on  $[a, b]$  is generated as  $X \sim F^{-1}[F(a) + U\{F(b) - F(a)\}]$  where  $U \sim U(0, 1)$  is standard uniform.

The function `samp_elif` also has an argument `upper` which serves for both right truncation, and right censoring. For the latter, any record simulated that exceeds `upper` is capped at that upper bound and declared partially observed. This is useful for simulating administrative censoring, whereby the birth date and the upper bound of the collection window fully determine whether an observation is right censored or not. An illustrative example is provided in the next section.

The `anova` method call uses the asymptotic null distribution for comparison of nested parametric distributions  $\mathcal{F}_0 \subseteq \mathcal{F}_1$ . We could use the bootstrap to see how good this approximation to the null distribution is. To mimic as closely as possible the data generating mechanism, which is custom in most scenarios, we condition on the sampling frame and the number of individuals in each birth cohort. The number dying at each age is random, but the right truncation limits will be the same for anyone in that cohort. We simulate excess lifetimes, then interval censor observations by keeping only the corresponding age bracket. Under the null hypothesis, the data are drawn from  $\hat{F}_0 \in \mathcal{F}_0$  and we generate observations from this right truncated distribution using the `samp_elif` utility, which also supports double interval truncation and left truncation right censoring. This must be done within a for loop since we have count attached to each upper bound, but the function is vectorized should we use a single vector containing all of the right truncation limits.

The bootstrap  $p$ -value for comparing models  $M_0 \subset M_1$  would be obtained by repeating the following steps  $B$  times and calculating the rank of the observed test statistic among alternatives:

1. Simulate new birth dates  $d_i$  ( $i = 1, \dots, n$ ) (e.g., drawing from a smoothed empirical distribution of birth dates); the latest possible birth date is one which ensures the person reached at least the threshold by the end of the period.
2. Subtract the endpoints of the sampling period, say  $c_1$  and  $c_2$  to get the minimum (maximum) age at death,  $c_1 - d_i$  (respectively  $c_2 - d_i$ ) days, which define the truncation bounds.
3. Use the function `samp_elif` to simulate new observations from a parametric interval truncated distribution from the null model  $M_0$
4. Use the optimization procedure in `fit_elif` to fit the model with both  $M_1$  and  $M_2$  and calculate the deviance, and from them the likelihood ratio statistic.

The algorithm is implemented below for comparing the Gompertz and the exponential model. Since the procedure is computationally intensive, users must trade off between the precision of the bootstrap  $p$ -value estimate and the number of replications,  $B$ .

```
set.seed(2022)
# Count of unique right truncation limit
db_rtrunc <- aggregate(count ~ rtrunc,
  FUN = "sum",
  data = japanese2,
  subset = age >= thresh)
B <- 1000L # Number of bootstrap replications
boot_anova <- numeric(length = B)
boot_gof <- numeric(length = B)
for(b in seq_len(B - 1L)){
  boot_samp <- # Generate bootstrap sample
  do.call(rbind, #merge data frames
    apply(db_rtrunc, 1, function(x){ # for each rtrunc and count
      count <- table( #tabulate count
        floor( #round down
          samp_elif( # sample right truncated exponential
            n = x["count"],
            scale = model0$par,
            family = "exp", #null model
            upper = x["rtrunc"] - thresh,
            type2 = "ltrt")))
      data.frame( # return data frame
        count = as.integer(count),
        rtrunc = as.numeric(x["rtrunc"]) - thresh,
        eage = as.integer(names(count)))
    })
  boot_mod0 <- # Fit null model to bootstrap sample
  with(boot_samp,
    fit_elif(time = eage,
      time2 = eage + 1L,
      rtrunc = rtrunc,
      type = "interval",
      event = 3,
      family = "exp",
      weights = count))
  boot_mod1 <- # Fit alternative model to bootstrap sample
  with(boot_samp,
    fit_elif(time = eage,
      time2 = eage + 1L,
      rtrunc = rtrunc,
      type = "interval",
```



```

    event = 3,
    family = "gomp",
    weights = count))
boot_anova[b] <- deviance(boot_mod0) -
  deviance(boot_mod1)
}
# Add original statistic
boot_anova[B] <- deviance(model1) - deviance(model0)
# Bootstrap p-value
(pval <- rank(boot_anova)[B] / B)
#> [1] 0.001

```

The asymptotic approximation is of similar magnitude as the bootstrap  $p$ -value. Both suggest that the more complex Gompertz model provides a significantly better fit.

## 2.4 Extreme value analysis

Extreme value theory suggests that, in many instances, the limiting conditional distribution of exceedances of a random variable  $Y$  with distribution function  $F$  is generalized Pareto, meaning

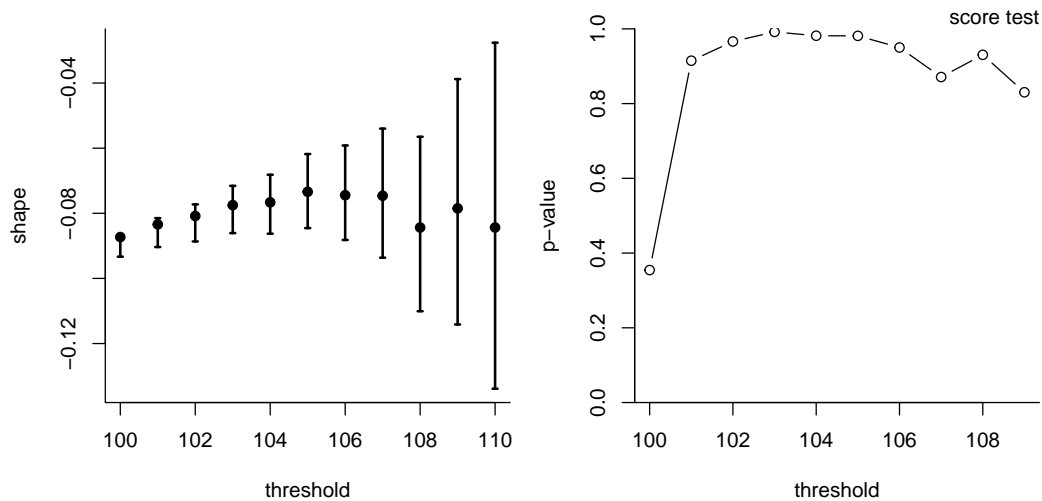
$$\lim_{u \rightarrow x^*} \Pr(Y - u > y \mid Y > u) = \begin{cases} (1 + \zeta y / \sigma)_+^{-1/\zeta}, & \zeta \neq 0; \\ \exp(-y / \sigma), & \zeta = 0; \end{cases} \quad (1)$$

with  $x_+ = \max\{x, 0\}$  and  $x^* = \sup\{x : F(x) < 1\}$ . This justifies the use of Equation (1) for the survival function of threshold exceedances when dealing with rare events. The model has two parameters: a scale  $\sigma$  and a shape  $\zeta$  which determines the behavior of the upper tail. Negative shape parameters correspond to bounded upper tails and a finite right endpoint for the support.

Study of population dynamics and mortality generally requires knowledge of the total population from which observations are drawn to derive rates. By contrast, the peaks over threshold method, by which one models the  $k$  largest observations of a sample, is a conditional analysis (e.g., given survival until a certain age), and is therefore free of denominator specification since we only model exceedances above a high threshold  $u$ . For modelling purposes, we need to pick a threshold  $u$  that is smaller than the upper endpoint  $x^*$  in order to have sufficient number of observations to estimate parameters. The threshold selection problem is a classical instance of bias-variance trade-off: the parameter estimators are possibly biased if the threshold is too low because the generalized Pareto approximation is not good enough, whereas choosing a larger threshold to ensure we are closer to the asymptotic regime leads to reduced sample size and increased parameter uncertainty.

To aid threshold selection, users commonly resort to parameter stability plots. These are common visual diagnostics consisting of a plot of estimates of the shape parameter  $\hat{\zeta}$  (with confidence or credible intervals) based on sample exceedances over a range of thresholds  $u_1, \dots, u_K$ . If the data were drawn from a generalized Pareto distribution, the conditional distribution above higher threshold  $v > u$  is also generalized Pareto with the same shape: this threshold stability property is the basis for extrapolation beyond the range of observed records. Indeed, if the change in estimation of  $\zeta$  is nearly constant, this provides reassurance that the approximation can be used for extrapolation. The only difference with survival data, relative to the classical setting, is that the likelihood must account for censoring and truncation. Note that, when we use threshold exceedances with a nonzero threshold (argument `thresh`), it however isn't possible to unambiguously determine whether left censored observations are still exceedances: such cases yield errors in the functions.

Theory on penultimate extremes suggests that, for finite levels and general distribution function  $F$  for which (1) holds, the shape parameter varies as a function of the threshold  $u$ , behaving like the derivative of the reciprocal hazard  $r(x) = \{1 - F(x)\} / f(x)$ . We can



**Figure 2:** Threshold diagnostic tools: parameter stability plots for the generalized Pareto model (left) and Northrop–Coleman  $p$ -value path (right) for the Japanese centenarian dataset. Both suggest that a threshold as low as 100 may be suitable for peaks-over-threshold analysis.

thus model the shape as piece-wise constant by fitting a piece-wise generalized Pareto model due to [Northrop and Coleman \(2014\)](#) and adapted in [Belzile et al. \(2022\)](#) for survival data. The latter can be viewed as a mixture of generalized Pareto over  $K$  disjoint intervals with continuity constraints to ensure a smooth hazard, which reduces to the generalized Pareto if we force the  $K + 1$  shape parameters to be equal. We can use a likelihood ratio test to compare the model, or a score test if the latter is too computationally intensive, and plot the  $p$ -values for each of the  $K$  thresholds, corresponding to the null hypotheses  $H_k : \xi_k = \dots = \xi_K$  ( $k = 1, \dots, K - 1$ ). As the model quickly becomes overparametrized, optimization is difficult and the score test may be a safer option as it only requires estimation of the null model of a single generalized Pareto over the whole range.

To illustrate these diagnostic tools, Figure 2 shows a threshold stability plot, which features a small increase in the shape parameters as the threshold increases, corresponding to a stabilization or even a slight decrease of the hazard at higher ages. We can envision a threshold of 108 years as being reasonable: the Northrop–Coleman diagnostic plot suggests lower thresholds are compatible with a constant shape above 100. Additional goodness-of-fit diagnostics are necessary to determine if the generalized Pareto model fits well.

```
par(mfrow = c(1, 2), mar = c(4, 4, 1, 1))
# Threshold sequence
u <- 100:110
# Threshold stability plot
tstab(arguments = args_japan,
       family = "gp",
       method = "profile",
       which.plot = "shape",
       thresh = u)
# Northrop-Coleman diagnostic based on score tests
nu <- length(u) - 1L
nc_score <- nc_test(arguments = c(args_japan, list(thresh = u)))
score_plot <- plot(nc_score)
graphics.off()
```

Each plot in the package can be produced using base R or using [ggplot2](#) ([Wickham, 2016](#)), which implements the grammar of graphics. To keep the list of package dependencies lean and adhere to the [tinyverse principle](#), the latter can be obtained by using the argument `plot.type` with the generic S3 method `plot`, or via `autoplot`, provided the [ggplot2](#) package is already installed.

## 2.5 Graphical goodness of fit diagnostics

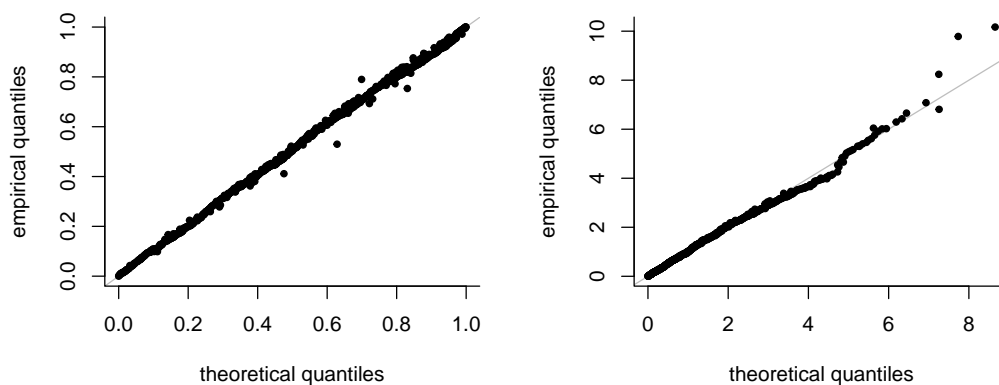
Determining whether a parametric model fits well to survival data is no easy task due to the difficulty in specifying the null distribution of many goodness of fit statistics, such as the Cramér–von Mises statistic, which differ for survival data. As such, the **longevity** package relies mostly on visual diagnostic tools. Waller and Turnbull (1992) discusses how classical visual graphics can be adapted in the presence of censoring. Most notably, only observed failure times are displayed on the  $y$ -axis against their empirical plotting position on the  $x$ -axis. Contrary to the independent and identically distributed case, the uniform plotting positions  $F_n(y_i)$  are based on the nonparametric maximum likelihood estimator discussed in Section 3.

The situation is more complicated with truncated data (Belzile et al., 2022), since the data are not identically distributed: indeed, the distribution function of observation  $Y_i$  truncated on the interval  $[a_i, b_i]$  is  $F_i(y_i) = \{F(y_i) - F(a_i)\} / \{F(b_i) - F(a_i)\}$ , so the data arise from different distributions even if these share common parameters. One way out of this conundrum is using the probability integral transform and the quantile transform to map observations to the uniform scale and back onto the data scale. Taking  $\tilde{F}(y_i) = F_n(y_i) = \text{rank}(y_i) / (n + 1)$  to denote the empirical distribution function estimator, a probability-probability plot would show  $x_i = \tilde{F}_i(y_i)$  against  $y_i = F_i(y_i)$ , leading to approximately uniform samples if the parametric distribution  $F$  is suitable. Another option is to standardize the observation, taking the collection  $\tilde{y}_i = F^{-1}\{F_i(y_i)\}$  of rescaled exceedances and comparing them to the usual plotting positions  $x_{(i)} = \{i / (n + 1)\}$ . The drawback of the latter approach is that the quantities displayed on the  $y$ -axis are not raw observations and the ranking of the empirical quantiles may change, a somewhat counter-intuitive feature. However, this means that the sample  $\{F_i(y_i)\}$  should be uniform under the null hypothesis, and this allows one to use methods from Säilynoja et al. (2022) to obtain point-wise and simultaneous confidence intervals.

**longevity** offers users the choice between three types of quantile-quantile plots: regular (Q-Q, "qq"), Tukey's mean detrended Q-Q plots ("tmd") and exponential Q-Q plots ("exp"). Other options on uniform scale are probability-probability (P-P, "pp") plots and empirically rescaled plots (ERP, "erp") (Waller and Turnbull, 1992), designed to ease interpretation with censored observations by rescaling axes. We illustrate the graphical tools with the dutch data above age 105 in Figure 3. The fit is correct above 110, but there is a notable dip due to excess mortality around 109.

```
fit_dutch <- fit_elife(
  arguments = dutch_data,
  event = 3,
  type = "interval2",
  family = "gp",
  thresh = 105,
  export = TRUE)
par(mfrow = c(1, 2))
plot(fit_dutch,
  which.plot = c("pp", "qq"))
```

Censored observations are used to compute the plotting positions, but are not displayed. As such, we cannot use graphical goodness of fit diagnostics for the Japanese interval censored data. An alternative, given that data are tabulated in a contingency table, is to use a chi-squared test for independence, conditioning on the number of individuals per birth cohort. The expected number in each cell (birth cohort and age band) can be obtained by computing the conditional probability of falling in that age band. The asymptotic null distribution should be  $\chi^2$  with  $(k - 1)(p - 1)$  degrees of freedom, where  $k$  is the number of age bands and  $p$  the number of birth cohorts. In finite samples, the expected count for large excess lifetimes are very low so one can expect the  $\chi^2$  approximation to be poor. To mitigate this, we can pool observations and resort to simulation to approximate the null



**Figure 3:** Probability-probability and quantile-quantile plots for generalized Pareto model fitted above age 105 years to Dutch data. The plots indicate broadly good agreement with the observation, except for some individuals who died age 109 for which too many have deaths close to their birthdates.

distribution of the test statistic. The bootstrap  $p$ -value for the exponential model above 108 years, pooling observations with excess lifetime of 5 years and above, is 0.872, indicating no evidence that the model is inadequate, but the test here may have low power.

## 2.6 Stratification

Demographers may suspect differences between individuals of different sex, from different countries or geographic areas, or by birth cohort. All of these are instances of categorical covariates. One possibility is to incorporate these covariates with suitable link function through parameters, but we consider instead stratification. We can split the data by levels of covariate (with factors) into sub-data and compare the goodness of fit of the  $K$  models relative to that which pools all observations. The `test_elif` function performs likelihood ratio tests for the comparisons. The test statistic is  $-2\{\ell(\hat{\theta}_0) - \ell(\hat{\theta})\}$ , where  $\hat{\theta}_0$  is the maximum likelihood estimator under the null model with common parameters, and  $\hat{\theta}$  is the unrestricted maximum likelihood estimator for the alternative model with the same distribution, but which allows for stratum-specific parameters. We illustrate this with a generalized Pareto model for the excess lifetime. The null hypothesis is  $H_0 : \sigma_f = \sigma_m, \zeta_f = \zeta_m$  against the alternative that at least one equality doesn't hold and so the hazard and endpoints are different.

```
print(
  test_elif(
    arguments = args_japan,
    thresh = 110,
    family = "gp",
    covariate = japanese2$gender)
)
#> Model: generalized Pareto distribution.
#> Threshold: 110
#> Number of exceedances per covariate level:
#> female   male
#>    642    61
#>
#> Likelihood ratio statistic: 0.364
#> Null distribution: chi-square (2)
#> Asymptotic p-value: 0.833
```

In the present example, there is no evidence against any difference in lifetime distribution between male and female; this is unsurprising given the large imbalance between counts for each covariate level, with far fewer males than females.

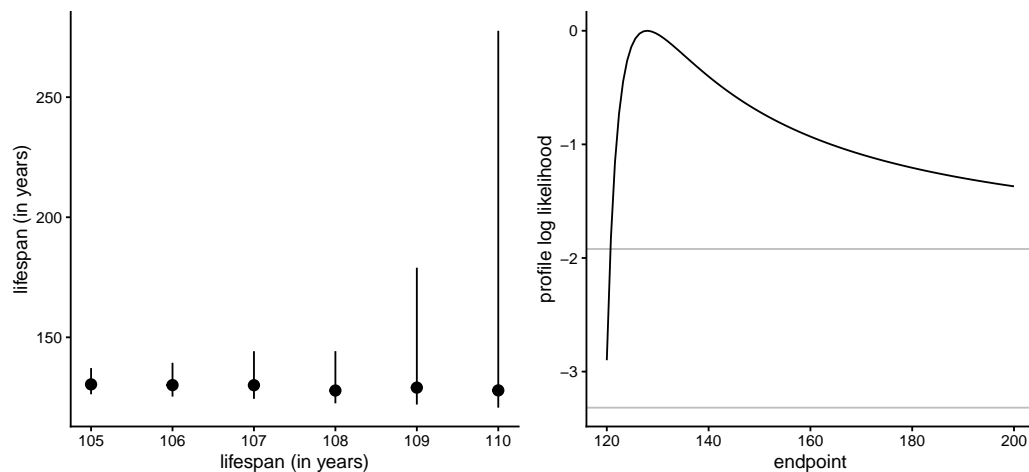
## 2.7 Extrapolation

If the maximum likelihood estimator of the shape  $\xi$  for the generalized Pareto model is negative, then the distribution has a finite upper endpoint; otherwise, the latter is infinite. With  $\xi < 0$ , we can look at the profile log likelihood for the endpoint  $\eta = -\sigma/\xi$ , using the function `prof_gp_endpt`, to draw the curve and obtain confidence intervals. The argument `psi` is used to give a grid of values over which to compute the profile log likelihood. The bounds of the  $(1 - \alpha)$  confidence intervals are obtained by fitting a cubic smoothing spline for  $y = \eta$  as a function of the shifted profile curve  $x = 2\{\ell_p(\eta) - \ell_p(\hat{\eta})\}$  on both sides of the maximum likelihood estimator and predicting the value of  $y$  when  $x = -\chi_1^2(1 - \alpha)$ . This technique works well unless the profile is nearly flat or the bounds lie beyond the range of values of `psi` provided; the user may wish to change them if they are too far. If  $\hat{\xi} \approx 0$ , then the upper bound of the confidence interval may be infinite and the profile log likelihood may never reach the cutoff value of the asymptotic  $\chi_1^2$  distribution.

The profile log likelihood curve for the endpoint, shifted vertically so that its value is zero at the maximum likelihood estimator, highlights the marked asymmetry of the distribution of  $\eta$ , shown in 4, with the horizontal dashed lines showing the limits for the 95% profile likelihood confidence intervals. These suggest that the endpoint, or a potential finite lifespan, could lie very much beyond observed records. The routine used to calculate the upper bound computes the cutoff value by fitting a smoothing spline with the role of the  $y$  and  $x$  axes reversed and by predicting the value of  $\eta$  at  $y = 0$ . In this example, the upper confidence limit is extrapolated from the model: more accurate measures can be obtained by specifying a longer and finer sequence of values of `psi` such that the profile log likelihood drops below the  $\chi_1^2$  quantile cutoff.

```
# Create grid of threshold values
thresholds <- 105:110
# Grid of values at which to evaluate profile
psi <- seq(120, 200, length.out = 101)
# Calculate the profile for the endpoint
# of the generalized Pareto at each threshold
endpt_tstab <- do.call(
  endpoint.tstab,
  args = c(
    args_japan,
    list(psi = psi,
         thresh = thresholds,
         plot = FALSE)))
# Compute corresponding confidence intervals
profile <- endpoint.profile(
  arguments = c(args_japan, list(thresh = 110, psi = psi)))
# Plot point estimates and confidence intervals
g1 <- autoplot(endpt_tstab, plot = FALSE, ylab = "lifespan (in years)")
# Plot the profile curve with cutoffs for conf. int. for 110
g2 <- autoplot(profile, plot = FALSE)
patchwork::wrap_plots(g1, g2)
```

Depending on the model, the conclusions about the risk of mortality change drastically: the Gompertz model implies an ever increasing hazard, but no finite endpoint for the distribution of exceedances. The exponential model implies a constant hazard and no endpoint. By contrast, the generalized Pareto can accommodate both finite and infinite endpoints. The marked asymmetry of the distribution of lifespan defined by the generalized



**Figure 4:** Maximum likelihood estimates with 95% confidence intervals as a function of threshold (left) and profile likelihood for exceedances above 110 years (right) for Japanese centenarian data. As the threshold increases, the number of exceedances decreases and the intervals for the upper bound become wider. At 110, the right endpoint of the interval would go until infinity.

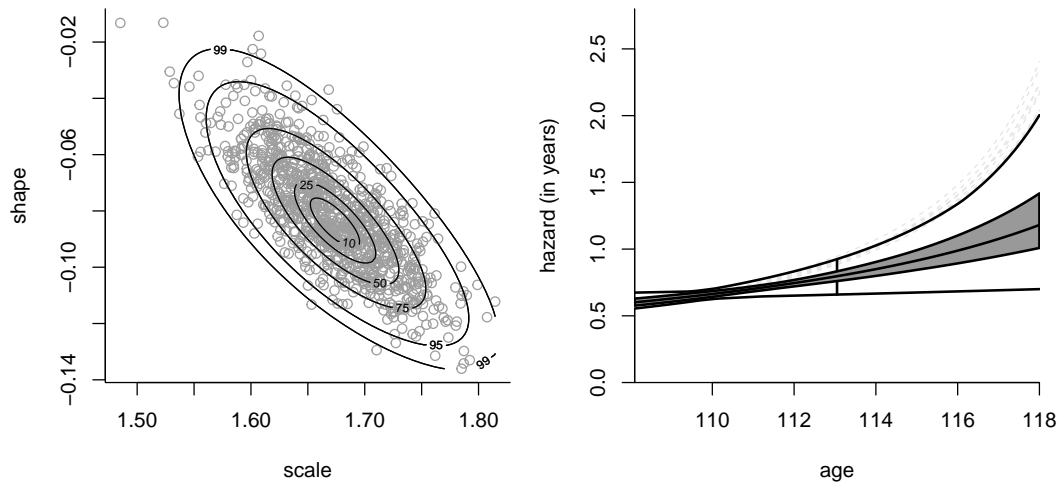
Pareto shows that inference obtained using symmetric confidence intervals (i.e., Wald-based) is likely very misleading: the drop in fit from having a zero or positive shape parameter  $\zeta$  is seemingly smaller than the cutoffs for a 95% confidence interval, suggesting that while the best point estimate is around 128 years, the upper bound is so large (and extrapolated) that everything is possible. The model however also suggests a very high probability of dying in any given year, regardless of whether the hazard is constant, decreasing or increasing.

## 2.8 Hazard

The parameters of the models are seldom of interest in themselves: rather, we may be interested in a summary such as the hazard function. At present, **longevity** does not allow general linear modelling of model parameters or time-varying covariates, but other software implementations can tackle this task. For example, **casebase** (Bhatnagar et al., 2022) fits flexible hazard models using logistic or multinomial regression with potential inclusion of penalties for the parameters associated to covariates and splines effects. Another alternative is **flexsurv** (Jackson, 2016), which offers 10 parametric models and allows for user-specified models. The **bshazard** package (Rebora et al., 2014) provides nonparametric smoothing via *B*-splines, whereas **muhas** handles kernel-based hazard for right censored data; both could be used for validation of the parametric models in the case of right censoring. The **rstpm2** package (Liu et al., 2018) handles generalized modelling for censoring with the Royston–Parmar model built from natural cubic splines (Royston and Parmar, 2002). Contrasting with all of the aforementioned approaches, we focus on parametric models: this is partly because there are few observations for the user case we consider and covariates, except perhaps for gender and birth year, are not available.

The hazard changes over time; the only notable exception is the exponential hazard, which is constant. **longevity** includes utilities for computing the hazard function from a fitted model object and computing point-wise confidence intervals using symmetric Wald intervals or the profile likelihood. Specifically, the `hazard_elif` function calculates the hazard  $h(t; \theta)$  point-wise at times  $t = x$ ; Wald-based confidence intervals are obtained using the delta-method, whereas profile likelihood intervals are obtained by reparametrizing the model in terms of  $h(t)$  for each time  $t$ . More naturally perhaps, we can consider a Bayesian analysis of the Japanese excess lifetime above 108 years. Using the likelihood and encoded log posterior provided in `logpost_elif`, we obtained independent samples from the posterior of the generalized Pareto parameters  $(\sigma, \zeta)$  with maximal data information prior using the **rust** package. Each parameter combination was then fed into `h_elif` and





**Figure 5:** Left: scatterplot of 1000 independent posterior samples from generalized Pareto model with maximum data information prior; the contour curves give the percentiles of credible intervals, and show approximate normality of the posterior. Right: functional boxplots for the corresponding hazard curves, with increasing width at higher ages.

the hazard evaluated over a range of values. Figure 5 shows the posterior samples and functional boxplots (Sun and Genton, 2011) of the hazard curves, obtained using the `fd` package. The risk of dying increases with age, but comes with substantial uncertainty, as evidenced by the increasing width of the boxes and interquartile range.

## 2.9 Nonparametric maximum likelihood estimation

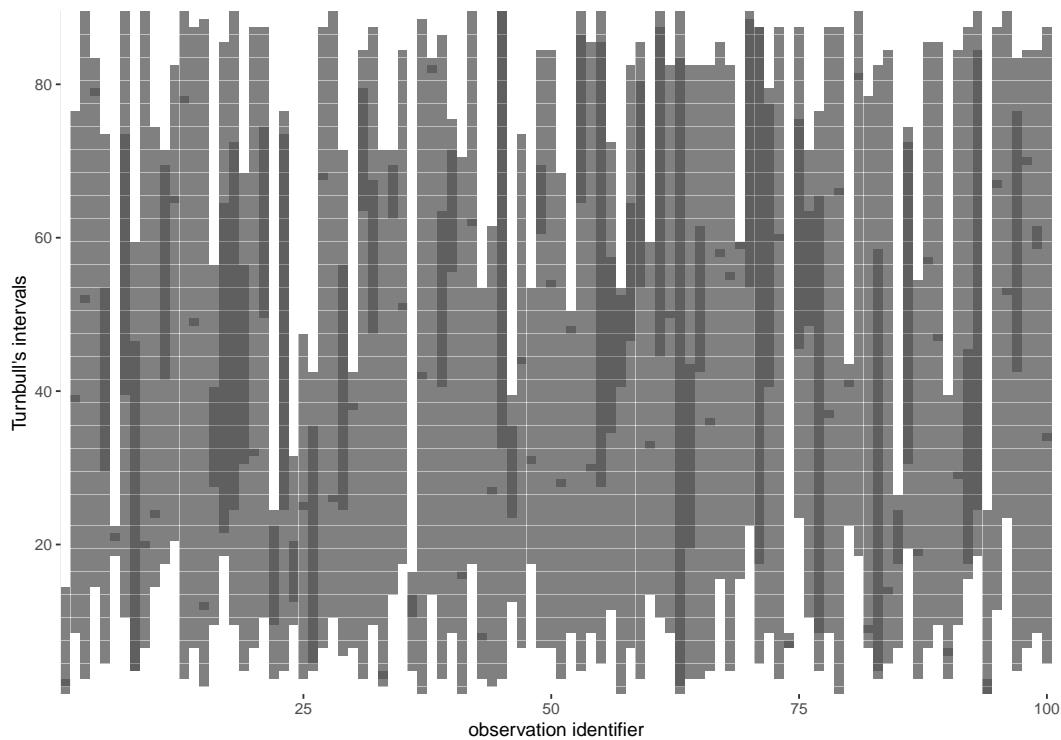
The nonparametric maximum likelihood estimator is unique only up to equivalence classes. The data for individual  $i$  consists of the tuple  $\{L_i, R_i, V_i, U_i\}$ , where the censoring interval is  $[L_i, R_i]$  and the truncation interval is  $[V_i, U_i]$ , with  $0 \leq V_i \leq L_i \leq R_i \leq U_i \leq \infty$ . Turnbull (1976) shows how one can build disjoint intervals  $C = \bigsqcup_{j=1}^m [a_j, b_j]$  where  $a_j \in \mathcal{L} = \{L_1, \dots, L_n\}$  and  $b_j \in \mathcal{R} = \{R_1, \dots, R_n\}$  satisfy  $a_1 \leq b_1 < \dots < a_m \leq b_m$  and the intervals  $[a_j, b_j]$  contain no other members of  $\mathcal{L}$  or  $\mathcal{R}$  except at the endpoints. This last condition notably ensures that the intervals created include all observed failure times as singleton sets in the absence of truncation. Other authors (Lindsey and Ryan, 1998) have taken interval censored data as semi-open intervals  $(L_i, R_i]$ , a convention we adopt here for numerical reasons. For interval censored and truncated data, Frydman (1994) shows that this construction must be amended by taking instead  $a_j \in \mathcal{L} \cup \{U_1, \dots, U_n\}$  and  $b_j \in \mathcal{R} \cup \{V_1, \dots, V_n\}$ .

We assign probability  $p_j = F(b_j^+) - F(a_j^-) \geq 0$  to each of the resulting  $m$  intervals under the constraint  $\sum_{j=1}^m p_j = 1$  and  $p_j \geq 0$  ( $j = 1, \dots, m$ ). The nonparametric maximum likelihood estimator of the distribution function  $F$  is then

$$\hat{F}(t) = \begin{cases} 0, & t < a_1; \\ \hat{p}_1 + \dots + \hat{p}_j, & b_j < t < a_{j+1} \quad (1 \leq j \leq m-1); \\ 1, & t > b_m; \end{cases}$$

and is undefined for  $t \in [a_j, b_j]$  ( $j = 1, \dots, m$ ).

The procedure of Turnbull can be encoded using  $m \times n$  matrices. For censoring, we build  $\mathbf{A}$  whose  $(i, j)$ th entry  $\alpha_{ij} = 1$  if  $[a_j, b_j] \subseteq A_i$  and zero otherwise. Since the intervals forming the set  $C$  are disjoint and in increasing order, a more storage efficient manner of keeping track of the intervals is to find the smallest integer  $j$  such that  $L_i \leq a_j$  and the largest  $R_i \geq b_j$  ( $j = 1, \dots, m$ ) for each observation. The same idea applies for the truncation sets  $B_i = (V_i, U_i)$  and matrix  $\mathbf{B}$  with  $(i, j)$  element  $\beta_{ij}$ .



**Figure 6:** Illustration of the truncation (pale grey) and censoring intervals (dark grey) equivalence classes based on Turnbull's algorithm. Observations must fall within equivalence classes defined by the former.

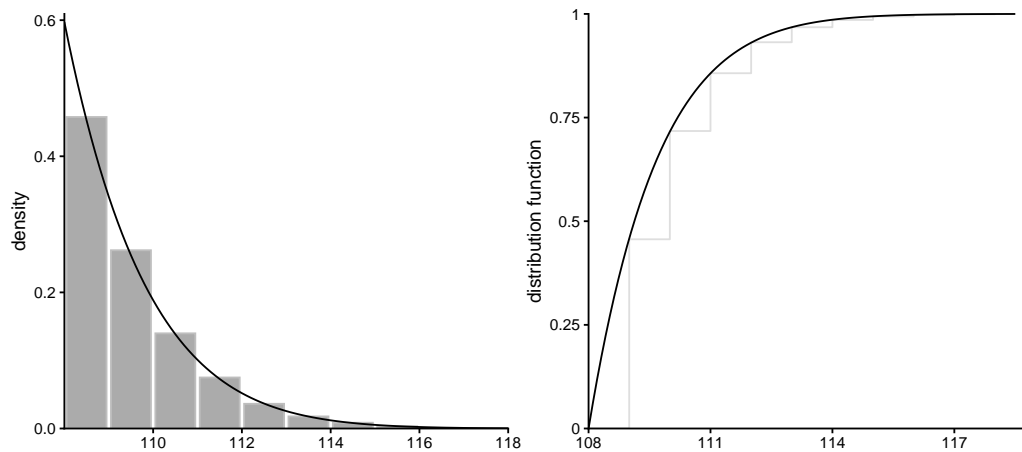
The log likelihood function is

$$\ell(\mathbf{p}) = \sum_{i=1}^n w_i \log \left( \sum_{j=1}^m \alpha_{ij} p_j \right) - w_i \log \left( \sum_{j=1}^m \beta_{ij} p_j \right)$$

The numerical implementation of the EM is in principle forward: first identify the equivalence classes  $C$ , next calculate the entries of  $A_i$  and  $B_i$  (or the vectors of ranges) and finally run the EM algorithm. In the second step, we need to account for potential ties in the presence of (interval) censoring and treat the intervals as open on the left for censored data. For concreteness, consider the toy example  $T = (1, 1, 2)$  and  $\delta = (0, 1, 1)$ , where  $\delta_i = 1$  if the observation is a failure time and  $\delta_i = 0$  in case of right censoring. The left and right censoring bounds are  $\mathcal{L} = \{1, 2\}$  and  $\mathcal{R} = \{1, 2, \infty\}$  with  $A_1 = (1, \infty)$ ,  $A_2 = \{1\}$  and  $A_3 = \{2\}$  and  $C_1 = \{1\}$ ,  $C_2 = \{2\}$ . If we were to treat instead  $A_1$  as a semi-closed interval  $[1, \infty)$ , direct maximization of the log likelihood in eq. 2.2 of [Turnbull \(1976\)](#) would give probability half to each observed failure time. By contrast, the Kaplan–Meier estimator, under the convention that right censored observations at time  $t$  were at risk up to and until  $t$ , assigns probability  $1/3$  to the first failure. To retrieve this solution with Turnbull's EM estimator, we need the convention that  $C_1 \notin A_1$ , but this requires comparing the bound with itself. The numerical tolerance in the implementation is taken to be the square root of the machine epsilon for doubles.

The maximum likelihood estimator (MLE) need not be unique, and the EM algorithm is only guaranteed a local maximum. For interval censored data, [Gentleman and Geyer \(1994\)](#) consider using the Karush–Kuhn–Tucker conditions to determine whether the probability in some intervals is exactly zero and whether the returned value is indeed the MLE.

Due to data scarcity, statistical inference for human lifespan is best conducted using parametric models supported by asymptotic theory, reserving nonparametric estimators to assess goodness of fit. The empirical cumulative hazard for Japanese is very close to linear



**Figure 7:** Nonparametric maximum likelihood estimate of the density (bar plot, left) and distribution function (staircase function, right), with superimposed generalized Pareto fit for excess lifetimes above 108 years. Except for the discreteness inherent to the nonparametric estimates, the two representations broadly agree at year marks.

from early ages, suggesting that the hazard may not be very far from exponential even if more complex models are likely to be favored given the large sample size.

The function `np_elif` returns a list with Turnbull's intervals  $[a_j, b_j]$  and the probability weights assigned to each, provided the latter are positive. It also contains an object of class `stepfun` with a weighting argument that defines a cumulative distribution function.

We can use the nonparametric maximum likelihood estimator of the distribution function to assess a fitted parametric model by comparing the density with a binned distribution, or the cumulative distribution function. The distribution function is defined over equivalence classes, which may be isolated observations or intervals. The data interval over which there is non-zero probability of events is broken down into equispaced bins and the probability of failing in the latter estimated nonparametrically based on the distribution function. Alternatively, users can also provide a set of breaks, or the number of bins.

```
ecdf <- np_elif(arguments = args_japan, thresh = 108)
# Summary statistics, accounting for censoring
round(summary(ecdf), digits = 2)
#>   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#> 109.00 110.00 111.00 110.01 111.00 118.00
# Plots of fitted parametric model and nonparametric CDF
model_gp <- fit_elif(
  arguments = args_japan,
  thresh = 108,
  family = "gp",
  export = TRUE)
# ggplot2 plots, wrapped to display side by side
patchwork::wrap_plots(
  autoplot(
    model_gp, # fitted model
    plot = FALSE, # return list of ggplots
    which.plot = c("dens", "cdf"),
    breaks = seq(0L, 8L, by = 1L) # set bins for histogram
  )
)
```

### 3 Conclusion

This paper describes the salient features of **longevity**, explaining the theoretical underpinning of the methods and the design considerations followed when writing the package. While **longevity** was conceived for modelling lifetimes, the package could be used for applications outside of demography. Survival data in extreme value theory is infrequent yet hardly absent. For example, rainfall observations can be viewed as rounded due to the limited instrumental precision of rain gauges and treated as interval-censored. Some historical records, which are often lower bounds on the real magnitude of natural catastrophes, can be added as right-censored observations in a peaks-over-threshold analysis. In insurance, losses incurred by the company due to liability claims may be right-censored if they exceed the policy cap and are covered by a reinsurance company, or rounded (Belzile and Nešlehová, 2025). In climate science, attribution studies often focus on data just after a record-breaking event, and the stopping rule leads to truncation (Miralles and Davison, 2023), which biases results if ignored.

The package has features that are interesting in their own right, including adapted quantile-quantile and other visual goodness of fit diagnostics for model validation. The testing procedures correctly handle tests for restrictions lying on the boundary of the parameter space. Parametric bootstrap procedures for such tests are not straightforward to implement given the heavy reliance on the data generating mechanism and the diversity of possible scenarios: this paper however shows how the utilities of the package can be coupled to ease such estimation and the code in the supplementary material illustrates how this can be extended for goodness of fit testing.

### Acknowledgements

The author thanks four anonymous reviewers for valuable feedback. This research was supported financially by the Natural Sciences and Engineering Research Council of Canada (NSERC) via Discovery Grant RGPIN-2022-05001.

### References

- J. J. Anderson. A vitality-based model relating stressors and environmental properties to organism survival. *Ecological Monographs*, 70(3):445–470, 2000. doi: 10.1890/0012-9615(2000)070[0445:avbmrs]2.0.co;2. URL [https://doi.org/10.1890/0012-9615\(2000\)070\[0445:AVBMRS\]2.0.CO;2](https://doi.org/10.1890/0012-9615(2000)070[0445:AVBMRS]2.0.CO;2). [p37]
- C. Anderson-Bergman. An efficient implementation of the EMICM algorithm for the interval censored NPMLE. *Journal of Computational and Graphical Statistics*, 26(2):463–467, 2017a. doi: 10.1080/10618600.2016.1208616. URL <https://doi.org/10.1080/10618600.2016.1208616>. [p38]
- C. Anderson-Bergman. icenReg: Regression models for interval censored data in R. *Journal of Statistical Software*, 81(12):1–23, 2017b. doi: 10.18637/jss.v081.i12. URL <https://doi.org/10.18637/jss.v081.i12>. [p38]
- E. Barbi, F. Lagona, M. Marsili, J. W. Vaupel, and K. W. Wachter. The plateau of human mortality: Demography of longevity pioneers. *Science*, 360(6396):1459–1461, 2018. ISSN 0036-8075. doi: 10.1126/science.aat3119. URL <https://doi.org/10.1126/science.aat3119>. [p43]
- R. E. Beard. A theory of mortality based on actuarial, biological and medical considerations. In *Proceedings of the International Population Conference, New York*, volume 1, London, 1963. International Union for the Scientific Study of Population. [p40]

- L. R. Belzile and J. G. Nešlehová. Statistics of extremes for incomplete data, with application to lifetime and liability claim modeling. In M. de Carvalho, R. Huser, P. Naveau, and B. J. Reich, editors, *Handbook on Statistics of Extremes*, chapter 31, page in press. CRC Press, 2025. [p54]
- L. R. Belzile, A. C. Davison, J. Gampe, H. Rootzén, and D. Zholud. Is there a cap on longevity? A statistical review. *Annual Review of Statistics and its Application*, 9:22–45, 2022. doi: 10.1146/annurev-statistics-040120-025426. URL <https://doi.org/10.1146/annurev-statistics-040120-025426>. [p38, 39, 41, 43, 46, 47]
- L. R. Belzile, C. Dutang, P. J. Northrop, and T. Opitz. A modeler’s guide to extreme value software. *Extremes*, 26:595–638, 2023. doi: 10.1007/s10687-023-00475-9. URL <https://doi.org/10.1007/s10687-023-00475-9>. [p37]
- S. R. Bhatnagar, M. Turgeon, J. Islam, J. A. Hanley, and O. Saarela. casebase: An alternative framework for survival analysis and comparison of event rates. *The R Journal*, 14:59–79, 2022. ISSN 2073-4859. doi: 10.32614/rj-2022-052. URL <https://doi.org/10.32614/RJ-2022-052>. [p50]
- C. G. Camarda. The curse of the plateau. measuring confidence in human mortality estimates at extreme ages. *Theoretical Population Biology*, 144:24–36, 2022. doi: 10.1016/j.tpb.2022.01.002. URL <https://doi.org/10.1016/j.tpb.2022.01.002>. [p42]
- H. Chernoff. On the distribution of the likelihood ratio. *The Annals of Mathematical Statistics*, 25(3):573–578, 1954. doi: 10.1214/aoms/1177728725. URL <https://doi.org/10.1214/aoms/1177728725>. [p42]
- M. L. Delignette-Muller and C. Dutang. fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, 64(4):1–34, 2015. doi: 10.18637/jss.v064.i04. [p37]
- L. Devroye. *Non-Uniform Random Variate Generation*. Springer, New York, 1986. URL <http://www.nrbook.com/devroye/>. [p43]
- B. Efron and V. Petrosian. Nonparametric methods for doubly truncated data. *Journal of the American Statistical Association*, 94(447):824–834, 1999. doi: 10.1080/01621459.1999.10474187. URL <https://doi.org/10.1080/01621459.1999.10474187>. [p38]
- J. J. Einmahl, J. H. J. Einmahl, and L. de Haan. Limits to human life span through extreme value theory. *Journal of the American Statistical Association*, 114(527):1075–1080, 2019. doi: 10.1080/01621459.2018.1537912. URL <https://doi.org/10.1080/01621459.2018.1537912>. [p38]
- M. P. Fay and P. A. Shaw. Exact and asymptotic weighted logrank tests for interval censored data: The interval R package. *Journal of Statistical Software*, 36(2):1–34, 2010. doi: 10.18637/jss.v036.i02. URL <https://doi.org/10.18637/jss.v036.i02>. [p38]
- H. Frydman. A note on nonparametric estimation of the distribution function from interval-censored and truncated observations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(1):71–74, 1994. ISSN 00359246. doi: 10.1111/j.2517-6161.1994.tb01960.x. URL <https://doi.org/10.1111/j.2517-6161.1994.tb01960.x>. [p51]
- R. Gentleman and C. J. Geyer. Maximum likelihood for interval censored data: Consistency and computation. *Biometrika*, 81(3):618–623, 1994. ISSN 0006-3444. doi: 10.1093/biomet/81.3.618. URL <https://doi.org/10.1093/biomet/81.3.618>. [p52]
- E. Gilleland and R. W. Katz. extRemes 2.0: An extreme value analysis package in R. *Journal of Statistical Software*, 72(8):1–39, 2016. doi: 10.18637/jss.v072.i08. [p37]
- P. Groeneboom and J. A. Wellner. *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Basel, Switzerland, 1992. ISBN 978-3-7643-2794-1. doi: 10.1007/978-3-0348-8621-5. URL <https://doi.org/10.1007/978-3-0348-8621-5>. [p38]

- C. Jackson. flexsurv: A platform for parametric survival modeling in R. *Journal of Statistical Software*, 70(8):1–33, 2016. doi: 10.18637/jss.v070.i08. URL <https://www.jstatsoft.org/index.php/jss/article/view/v070i08>. [p50]
- D. A. Jdanov, V. M. Shkolnikov, and S. Gellers-Barkmann. The international database on longevity: Data resource profile. In H. Maier, B. Jeune, and J. W. Vaupel, editors, *Exceptional Lifespans*, Demographic Research Monographs, pages 22–24. Springer, Cham, Switzerland, 2021. doi: 10.1007/978-3-030-49970-9\_2. URL [https://doi.org/10.1007/978-3-030-49970-9\\_2](https://doi.org/10.1007/978-3-030-49970-9_2). [p43]
- E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958. doi: 10.1080/01621459.1958.10501452. URL <https://doi.org/10.1080/01621459.1958.10501452>. [p38]
- T. Li and J. J. Anderson. The vitality model: A way to understand population survival and demographic heterogeneity. *Theoretical Population Biology*, 76(2):118–131, 2009. ISSN 0040-5809. doi: 10.1016/j.tpb.2009.05.004. URL <https://doi.org/10.1016/j.tpb.2009.05.004>. [p37]
- J. C. Lindsey and L. M. Ryan. Methods for interval-censored data. *Statistics in Medicine*, 17(2):219–238, 1998. doi: 10.1002/(sici)1097-0258(19980130)17:2<219::aid-sim735>3.0.co;2-o. URL [https://doi.org/10.1002/\(SICI\)1097-0258\(19980130\)17:2<219::AID-SIM735>3.0.CO;2-O](https://doi.org/10.1002/(SICI)1097-0258(19980130)17:2<219::AID-SIM735>3.0.CO;2-O). [p51]
- X.-R. Liu, Y. Pawitan, and M. Clements. Parametric and penalized generalized survival models. *Statistical Methods in Medical Research*, 27(5):1531–1546, 2018. doi: 10.1177/0962280216664760. URL <https://doi.org/10.1177/0962280216664760>. [p50]
- H. Maier, B. Jeune, and J. W. Vaupel, editors. *Exceptional Lifespans*. Demographic Research Monographs. Springer, 2021. doi: 10.1007/978-3-030-49970-9. URL <https://doi.org/10.1007/978-3-030-49970-9>. [p38]
- O. Miralles and A. C. Davison. Timing and spatial selection bias in rapid extreme event attribution. *Weather and Climate Extremes*, 41:100584, 2023. doi: 10.1016/j.wace.2023.100584. URL <https://doi.org/10.1016/j.wace.2023.100584>. [p54]
- P. J. Northrop and C. L. Coleman. Improved diagnostic plots for extreme value analyses. *Extremes*, 17:289–303, 2014. doi: 10.1007/s10687-014-0183-z. URL <https://doi.org/10.1007/s10687-014-0183-z>. [p46]
- W. Perks. On some experiments in the graduation of mortality statistics. *Journal of the Institute of Actuaries*, 63(1):12–57, 1932. doi: 10.1017/s0020268100046680. [p40]
- P. Rebora, A. Salim, and M. Reilly. bshazard: A flexible tool for nonparametric smoothing of the hazard function. *The R Journal*, 6(2):114–122, 2014. doi: 10.32614/rj-2014-028. URL <https://doi.org/10.32614/RJ-2014-028>. [p50]
- S. J. Richards. A handbook of parametric survival models for actuarial use. *Scandinavian Actuarial Journal*, 2012(4):233–257, 2012. doi: 10.1080/03461238.2010.506688. URL <https://doi.org/10.1080/03461238.2010.506688>. [p40]
- A. Rotnitzky, D. R. Cox, M. Bottai, and J. Robins. Likelihood-based inference with singular information matrix. *Bernoulli*, 6(2):243 – 284, 2000. doi: bj/1081788028. URL <https://doi.org/>. [p42]
- P. Royston and M. K. B. Parmar. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21(15):2175–2197, 2002. doi: 10.1002/sim.1203. URL <https://doi.org/10.1002/sim.1203>. [p50]



- T. Säilynoja, P.-C. Bürkner, and A. Vehtari. Graphical test for discrete uniformity and its applications in goodness of fit evaluation and multiple sample comparison. *Statistics and Computing*, 32(2):32, 2022. doi: 10.1007/s11222-022-10090-6. URL <https://doi.org/10.1007/s11222-022-10090-6>. [p47]
- S. G. Self and K.-Y. Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610, 1987. doi: 10.1080/01621459.1987.10478472. URL <https://doi.org/10.1080/01621459.1987.10478472>. [p42]
- H. Ševčíková, N. Li, V. Kantorová, P. Gerland, and A. E. Raftery. *Age-Specific Mortality and Fertility Rates for Probabilistic Population Projections*, pages 285–310. Springer, Cham, Switzerland, 2016. ISBN 978-3-319-26603-9. doi: 10.1007/978-3-319-26603-9\_15. URL [https://doi.org/10.1007/978-3-319-26603-9\\_15](https://doi.org/10.1007/978-3-319-26603-9_15). [p37]
- P.-S. Shen. Nonparametric analysis of doubly truncated data. *Annals of the Institute of Statistical Mathematics*, 62(5):835–853, 2010. ISSN 1572-9052. doi: 10.1007/s10463-008-0192-2. URL <https://doi.org/10.1007/s10463-008-0192-2>. [p38, 43]
- A. G. Stephenson. evd: Extreme value distributions. *R News*, 2(2), 2002. URL [https://cran.r-project.org/doc/Rnews/Rnews\\_2002-2.pdf](https://cran.r-project.org/doc/Rnews/Rnews_2002-2.pdf). [p37]
- Y. Sun and M. G. Genton. Functional boxplots. *Journal of Computational and Graphical Statistics*, 20(2):316–334, 2011. doi: 10.1198/jcgs.2011.09224. [p51]
- T. M. Therneau and P. M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, New York, 2000. ISBN 0-387-98784-3. doi: 10.1007/978-1-4757-3294-8. URL <https://doi.org/10.1007/978-1-4757-3294-8>. [p37, 39]
- W.-Y. Tsai, N. P. Jewell, and M.-C. Wang. A note on the product-limit estimator under right censoring and left truncation. *Biometrika*, 74(4):883–886, 1987. doi: 10.1093/biomet/74.4.883. URL <https://doi.org/10.1093/biomet/74.4.883>. [p38]
- B. W. Turnbull. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B*, 38:290–295, 1976. doi: 10.1111/j.2517-6161.1976.tb01597.x. URL <https://doi.org/10.1111/j.2517-6161.1976.tb01597.x>. [p38, 51, 52]
- L. A. Waller and B. W. Turnbull. Probability plotting with censored data. *American Statistician*, 46(1):5–12, 1992. doi: 10.1080/00031305.1992.10475837. URL <https://doi.org/10.1080/00031305.1992.10475837>. [p47]
- S. Watanabe. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594, 2010. ISSN 1532-4435. URL <https://dl.acm.org/doi/10.5555/1756006.1953045>. [p42]
- H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL <https://ggplot2.tidyverse.org>. [p46]
- Y. Ye. *Interior Algorithms for Linear, Quadratic, and Linearly Constrained Non-Linear Programming*. PhD thesis, Department of ESS, Stanford University, 1987. [p38]
- T. W. Yee. *Vector Generalized Linear and Additive Models*. Springer, 2015. doi: 10.1007/978-1-4939-2818-7. URL <https://doi.org/10.1007/978-1-4939-2818-7>. [p38]
- T. W. Yee and C. J. Wild. Vector generalized additive models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(3):481–493, 1996. doi: 10.1111/j.2517-6161.1996.tb02095.x. URL <https://doi.org/10.1111/j.2517-6161.1996.tb02095.x>. [p38]
- B. D. Youngman. evgam: An R package for generalized additive extreme value models. *Journal of Statistical Software*, 103(3):1–26, 2022. doi: 10.18637/jss.v103.i03. URL <https://www.jstatsoft.org/index.php/jss/article/view/v103i03>. [p37]

*Léo R. Belzile*  
*Department of Decision Sciences, HEC Montréal*  
*3000, chemin de la Côte-Sainte-Catherine*  
*Montréal (Québec), Canada*  
*H3T 2A7*  
<https://lbelzile.bitbucket.io>  
*ORCID: 0000-0002-9135-014X*  
[leo.belzile@hec.ca](mailto:leo.belzile@hec.ca)