UNIVERSITEIT VAN AMSTERDAM

Amsterdam School of Economics

Computer class NLIST—Assignment 1B

Makeham and Gompertz survival distributions

The objective of this exercise is to get acquainted with

- cdf F, survival function $S = 1 - F$, density F' and mortality rate $\mu(x) = -(\log S(x))'$ of Makeham and Gompertz distributions for life-times / failure times
- generating samples from these distributions
- computing the expected value of the lifetime by integrating $S(x)$
- the effect on this of changing the parameters by a little
- finding the (log)likelihood of a sample
- using the `optim` routine to find the ML-estimates of the parameters
- comparing the Makeham fit for a sample to a Gompertz fit, using AIC and BIC
- making plots of survival functions and mortality rates
- finding Makeham parameters when only the numbers of deaths and the exposures at integer ages are given, rather than lengths of lifetimes

1 Survival distributions and hazard / mortality rates

The mortality rate of a continuous random variable X denoting a life-time is

$$\mu_X(x) \stackrel{\text{def}}{=} \frac{f_X(x)}{1 - F_X(x)}, \quad x > 0. \tag{1}$$

So the conditional probability of death between x and $x + \mathrm{d}x$ for someone alive at age x is equal to

$$\frac{F_X(x + \mathrm{d}x) - F_X(x)}{1 - F_X(x)} = \frac{f_X(x)\,\mathrm{d}x}{1 - F_X(x)} = \mu_X(x)\,\mathrm{d}x.$$

Life-time X has a Makeham(a, b, c) distribution if $\mu_X(x) = a + bc^x$. Gompertz' law is the special case with $a = 0$. In our parameterization, we require $a \geq 0$, $b \geq 0$, $c \geq 1$. Other parameterizations can also be found in the literature, like $\mu_X(x) = a_2 + a_1\,\mathrm{e}^{x/b}$ as in R's `eha` package, or $\mu_X(x) = \alpha\,\mathrm{e}^{\beta x} + \lambda$ as in Wikipedia.

To find $S(x) = \Pr[X > x]$ for these distributions, use that $\mu_X(x) = -\frac{\mathrm{d}}{\mathrm{d}x} \log S(x)$ by the chain rule and (1), so from $\mu_X(x) = a + bc^x$ we get

$$-\log(S(x)) = ax + \frac{b}{\log c} c^x + C \text{ for some constant } C.$$

From $S(0) = 1$ we find $C = -b/\log c$, and therefore

$$S(x) = 1 - F_X(x) = \exp\left(-ax - \frac{b}{\log c}(c^x - 1)\right), \quad x > 0. \tag{2}$$

Some remarks on mortality rates/laws:

- As well as density or mgf, the mortality rate, aka the *failure rate*, characterizes a life-time distribution.
- From (1), we find $f_X(x) = \mu_X(x)S_X(x)$.
- If $X = \min(Y, Z)$ and Y, Z are independent, then $1 - F_X(x) = \Pr[Y > x]\Pr[Z > x]$, hence $\mu_X(x) = \mu_Y(x) + \mu_Z(x)$.
- So Makeham's mortality law $\mu_X(x) = a + bc^x$ arises from a *competing risks* model, where people die at age X from the first of these causes:
  - 'accident' at age $Z \sim$ exponential(a) (note that in that case it can easily be shown that $\mu_Z(x) \equiv a$), or
  - 'senescence' at age $Y \sim$ Gompertz(b, c).

Q_1 Let $Y = \log(1 + \frac{V \log c}{b})/\log c$, with $V \sim$ exponential(1).

Prove that $c^Y - 1 = \frac{V \log c}{b}$, and therefore $c^Y - 1 \sim$ exponential$(\frac{b}{\log c})$.

Then find $\Pr[Y > x] = \Pr[c^Y - 1 > c^x - 1]$, and compare with (2) to see that $Y \sim$ Gompertz(b, c). □

Using this result, we will define a function to draw a sample of size n from a Makeham(a, b, c) distribution, including its special case the Gompertz(b, c) distribution with $a = 0$. Following the competing risks model, it returns $X = \min(Y, Z)$ with $Y \sim$ Gompertz(b, c) and $Z \sim$ exponential(a). Note the use of the `pmin` function to get the *parallel* minima of two vectors.

```
gen.Sample <- function(n, a, b, c)
{if (any(a<0,b<0,c<1)) stop("Invalid parameters")
 lifetimes <- log(1+rexp(n)*log(c)/b)/log(c)
 if (a>0) lifetimes <- pmin(lifetimes, rexp(n)/a)
 return(lifetimes)}

set.seed(2525); G <- gen.Sample(2000, 0, 8e-5, 1.08)
set.seed(2525); M <- gen.Sample(2000, 5e-4, 8e-5, 1.08)
all(G>=M)  ## TRUE
mean(M==G) ## 0.959
rbind(Gompertz=summary(G), Makeham=summary(M))
#            Min. 1st Qu. Median  Mean 3rd Qu.  Max.
# Gompertz 4.9450   73.72  85.12 81.98   93.73 116.6
# Makeham  0.2809   71.85  84.15 80.08   93.35 116.6
```

In the last part, we first generate a Gompertz sample `G`. Resetting the random number generator, we generate a Makeham sample `M`, with the same b, c but $a > 0$. From the function definition we see that the Makeham lifetimes arise as the minimum of the Gompertz lifetimes and exponential random variables.

In the last line, we compare both samples regarding the important statistics quartiles and mean.

Q_2 Why are all summary statistics given smaller for Makeham?
What can you tell about the probability of dying by old age rather than accident with these parameters for the lifetime distribution? □

Note that the life-times produced here are reals, not truncated to integers (kurtate life-times).

2 Mean life-times

In (2), we derived an expression for the Makeham survival function S with parameters a, b, c. The following is a simple `R`-function to compute it:

```
S <- function(x, a, b, c) exp(-a*x - b/log(c)*(c^x-1))
```

Since $\mathrm{E}[X] = \int_0^\infty S(x)\,\mathrm{d}x$ by MART (1.33) at $d = 0$, we can calculate the theoretical expected value of a lifetime and compare it to the mean of a large sample, as follows (note the a, b, c parameters being passed on as additional arguments `...` (ellipsis) to the `S` function; see `?integrate`):

```
a <- 5e-4; b <- 8e-5; c <- 1.09
mean.age <- integrate(S, 0, Inf, a, b, c)$value  ##72.98
mean(gen.Sample(1e6, a, b, c))  ## 73.00
```

Q_3 We want to find out about the effect on the theoretical mean life-time of changes in each of the parameters. For this, calculate the mean life-times at (a', b, c), (a, b', c) and (a, b, c'), divided by `mean.age`, when $a' = 1.02a$, $b' = 1.02b$, $c' - 1 = 1.02(c - 1)$.
Which change affects the mean life-time the most? □

3 ML estimation of the Makeham parameters

The function below computes the log of the density $f_X(x) = \mu_X(x)S_X(x)$, see (2):

```
log.fx <- function(x, a, b, c) log(a + b*c^x) + (-a*x - b/log(c)*(c^x-1))
```

Q_4 Apart from mathematical convenience because of multiplications reducing to additions, what other reason is there to look at logarithms of the likelihood here? Note that we want to compute the maximum likelihood of the parameters a, b, c based on a large sample with a likelihood of about e^{-8300}. □

The optimization routine `optim()` must be fed a function of a parameter vector. The likelihood is the product of all density values, so the loglikelihood is the sum of the logs of the density values. Since `optim()` minimizes its function in standard mode, to maximize the loglikelihood of the Makeham parameters for the `M` sample created earlier, we insert a minus sign into the loglikelihood function:

```
log.Lik <- function(p) {- sum(log.fx(M, p[1], p[2], p[3]))}
```

We store the object resulting from the optimization in `o`, and print the function value at the optimum and the optimal parameter values (think 'argmax'), like this:

```
a <- 5e-4; b <- 8e-5; c <- 1.08
o <- optim(c(a,b,c), log.Lik) ## 26 warnings "In log(a+b*c^x): NaNs produced"
o$value  ##  8428.489
o$par    ##  a=7.3e-04 b=6.0e-05 c=1.083
```

The first parameter of `optim()` is the initial value. We had a good idea where the maximizing parameter values might be located. The choice of the initial values might make a lot of difference. Note that your initial choice of parameters should be such that sensible values for estimated probabilities of dying result over the whole age range $x = 0, \ldots, 110$. At $x = 100$, the survival probability $S(x) = \exp(-ax - \frac{b}{\log c}(c^x - 1))$ should be somewhere around one half. If you get inexplicable results, try inspecting $S(100)$ and $S(10)$ for your choices of the initial parameters.

To compute the maximizing b, c for the Gompertz density on the `M` sample, we must replace the function to be optimized by the following:

```
log.Lik <- function(p) {-sum(log.fx(M, 0, p[1], p[2]))} ## Gompertz on M
o <- optim(c(b,c), log.Lik)   ## 13x NaN
o$value  ##  8456.258
o$par    ##  a=0 b=1.3e-4 c=1.074
```

Q_5 Make the appropriate changes to compute ML-estimates of (`a`,) `b`, `c` for the `G` sample, using (i) the Makeham density and (ii) the Gompertz density.

Hint: The function to be optimized has a minus-sign in it, because `optim()` in its standard mode minimizes. The resulting optimal values `lAB` for *the negative of* the loglikelihood using the `A` density on the `B` sample should be:

```
lMM = 8428.489;  lGM = 8456.258;  lMG = 8245.877; lGG = 8248.117
```

□

4 Comparing the fits by using AIC and BIC

Akaike's Information Criterion (AIC) is defined as $-2\log L + 2k$ [see MART (9.33)], with L the maximized likelihood and k the number of parameters estimated.

Q_6 For the Makeham sample, which model is better as regards AIC: Makeham or Gompertz? Same question for the Gompertz sample.
In BIC (9.34), the penalty for parameters is $k \log n \approx 7.6k$ in our case. Now which models are better? □

5 Graphical illustrations

The following plots depict the cdf $F(x)$ of the lifetime for various values of the parameters a, b, c:

```
S <- function(x, a, b, c) exp(-a*x - b/log(c)*(c^x-1))
x <- 0:110
plot(x, 1-S(x, 0, 8e-5, 1.08), type="h", ylab="F(x; a,b,c)")
lines(x, 1-S(x, 1e-3, 8e-5, 1.08), col="red")
points(x, 1-S(x, 0, 8e-5, 1.09), col="blue")
```

Q_7 Describe the effects on the cdf of varying a or c. □

Q_8 Make a similar plot for the mortality rate μ. □

6 Counts instead of continuous lifetimes

Up to now in this assignment, we have looked at data involving the exact lifetimes of people. The data found in practice, however, involve counts of people alive and dying by age group. Such data can be downloaded from sites like the Dutch Central Bureau of Statistics (CBS) or from the Human Mortality Database. These data are transformed in an attempt to represent as well as possible the numbers of people dying at certain integer ages as well as the number exposed to that risk in a calendar year, resulting in fractional numbers. We ignore such niceties for the purpose of this exercise. We assume that, after rounding, the numbers published simply represent the number of people of a certain age dying in a year, as well as the number of people having that age at the beginning of the year. From these data, we want to estimate the parameters of an underlying Makeham distribution of lifetimes.

By running the following script you get mortality data about the Dutch population over a period of 58 years. In D_{xt} the number of deaths is recorded on ages $\in (x-1, x]$, $x = 1, \ldots, 101$ in period $t = 1, \ldots, 58$, in e_{xt} the number of people in that age group. Ordinarily, D_{xt} denotes the number of people with age in $[x, x+1)$ that die, but since indices in `R` start with 1, not 0, we use this notation. Though we now have the data to investigate the evolution of

mortality probabilities over time, at the moment we are not interested in this, so we just look at the *total* numbers of people dying $D_x = \sum_t D_{xt}$ and alive $e_x = \sum_t e_{xt}$ at an age in $(x-1, x]$, $x = 1, \ldots, 101$. The Human Mortality Database provides data like this, but we have downloaded these data and put them in directly readable `.csv`-files for your convenience.

```
path <- "http://www1.fee.uva.nl/ke/act/people/kaas/"
D.xt <- round(scan(paste(path,"deaths.csv",sep=""), sep=";", dec=","))
e.xt <- round(scan(paste(path,"exposures.csv",sep=""), sep=";", dec=","))
nages <- 101; nyears <- 58
D.xt <- matrix(D.xt,nages,nyears,byrow=TRUE); D.x <- apply(D.xt,1,sum)
e.xt <- matrix(e.xt,nages,nyears,byrow=TRUE); e.x <- apply(e.xt,1,sum)
```

By construction, `D.x[1:101]` approximates the number of deaths between ages $x-1$ and x, totaled over t, and `e.x[1:101]` the total of the number of people alive of that age; $x = 1, \ldots, 101$.

Let q_x, to be stored in `q.x[1:101]`, be the probability of dying between $x-1$ and x, given survival to $x-1$. Note that q_x is the one-year equivalent of the infinitesimal mortality rate $\mu_X(x)$, see (1), in the continuous case. We will treat the values `D.x[1:101]` as realizations of Binomial(`e.x[x], q.x[x]`) random variables. Obviously people of integer age $x-1$ years might actually be anywhere from $x-1$ to x years old, but to compute q_x here we assume that they are exactly $x-1$, and if $X \sim S()$ denotes the lifetime, then

$$q_x = \Pr[x-1 < X \leq x \,|\, x-1 < X] = \frac{S(x-1) - S(x)}{S(x-1)}.$$

Q_9 Assuming that the lifetimes obey a Makeham(a, b, c) law, construct the vector `q.x`. For this, compute the survival function (aka the decumulative distribution function) $S(x; a, b, c)$ in arguments $x = 0, \ldots, 101$ and apply the `diff()` and the `head()` functions to the resulting vector. Fill in the dots below:

```
S <- function(x, a, b, c) exp(-a*x - b/log(c)*(c^x-1))
agerange <- 1:nages  ## denotes the range of ages accounted for when finding ML
LogLik <- function(p)
  { ddfs <- S(0:nages, p[1], p[2], p[3])
    q.x <- ...
    -sum(dbinom(D.x[agerange], e.x[agerange], q.x[agerange], log=TRUE))}
```

The last line computes the negative of the loglikelihood of the a, b, c parameter values, given the data, that is, `D.x[x]` 'successes' in `e.x[x]` trials with probability of 'success' `e.x[x]`, `x` $\in$ `agerange`.

Ignoring warnings when invalid parameters values are tried, find the ML-estimates for a, b, c based on this sample.

Plot both the optimally estimated $\widehat{q_x}$ values and the recorded fractions D_x/e_x of people dying. You see that at ages in $(0, 1]$, Makeham's Law is not valid. Also, the observed values at the end are lower than the fitted values, ostensibly because the parameter estimates tried to fit the value at $x = 1$, distorting the fit. So for the estimation of the parameter values,

leave out $x = 1$, considering only age range `agerange <- 2:nages`. Estimate the parameters again, and make the same plots. You should observe that the fit for high ages is better.

In these plots, it is hardly visible that in fact for ages around $x = 20$, mortality is much higher than predicted by Makeham. To be able to see this better, compare `q.x[1:60]` with `(D.x/e.x)[1:60]`. This extra mortality around age 20 is generally referred to as the 'accident hump'. □