Predictive Analytics for Modeling Threshold Life Tables

Min Ji, Mostafa Aminzadeh, Min Deng

Towson University, Maryland, USA

Advances in Predictive Analytics Conference
Waterloo , December 2, 2017

Outline

The generalized Pareto distribution

- The peaks-over-threshold (POT) approach in the extreme value theory describes the behavior of large observations which exceed some high threshold;
- The generalized Pareto distribution (GPD) provides a unifying approach to the distribution of threshold excesses;
- In actuarial science literature, the (GPD) has been regarded as a promising approach to the modeling of high age mortality;

Excess distribution

For a large value u, the conditional random variable over the threshold u is defined to be $Y = X - u | X > u$, with the excess distribution expressed as

$$F_u(y) = \Pr\{X - u \leq y | X > u\} = \frac{F(u+y) - F(u)}{1 - F(u)}, \qquad (1)$$

for $0 \leq y < x_F - u$, where $x_F < \infty$ is the right endpoint of F if it exists.

The general Pareto distribution

The excess distribution $F_u(y)$ over some suitably high threshold u can be well approximated by a generalized Pareto distribution given by

$$G(y) = \begin{cases} 1 - \left(1 + \frac{\gamma y}{\sigma}\right)^{-1/\gamma}, & \text{if } \gamma \neq 0 \\ 1 - \exp\left(-\frac{y}{\sigma}\right), & \text{if } \gamma = 0 \end{cases} \quad (2)$$

defined on $\{y : y > 0 \text{ and } (1 + \gamma y/\sigma) > 0\}$. If $\gamma < 0$ the distribution of threshold excesses has an upper bound of $-\sigma/\gamma$; if $\gamma \geq 0$ the distribution has no upper limit. The case of $\gamma = 0$ is interpreted by taking the limit $\gamma \to 0$, leading to an exponential distribution with parameter $1/\sigma$.

Research question

- A threshold life table has been proposed by Li et al. (2008), using the Gompertz law connected with the GPD;
- There are difficulties in parameter estimation for the threshold life tables;
- Some research used a predetermined threshold;
- Thresholds were set at integer ages;
- Cannot guarantee smooth death probabilities over ages.

Research objectives

- To achieve a smooth threshold life table
- To use Bayesian approach to estimate the parameters
- To derive predictive density of lifetime distribution and other quantities of interest

Some notations

Define the following notations for life time distribution

- X: the age at death random variable, assumed to be continuous.
- $f(x)$: the probability density function of X.
- $F(x)$: the distribution function of X.
- $S(x) = 1 - F(x)$: the survival function.
- $\mu(x) = f(x)/S(x)$: the force of mortality,which is called the failure rate in credibility theory.
- $d(x)$: the number of death at age x.
- $E(x)$: the amount of exposure-to-risk at age x.

Threshold life tables

The threshold life table proposed by Li et al. (2008) is mathematically expressed as follows:

$$F(x) = \begin{cases} 1 - \exp\left(-\frac{B}{\ln c}(c^x - 1)\right) & \text{if } x \le u \\ F(u) + (1 - F(u))\left(1 - \left(1 + \frac{\gamma(x-u)}{\sigma}\right)^{-1/\gamma}\right) & \text{if } x > u \end{cases}, \tag{3}$$

where $F(x)$ is the distribution function of the lifetime random variable.

Propose two constraints

1. To ensure $F(x)$ is continuous and differentiable at $x = u$, we need
$$\sigma = \frac{1}{Bc^u},$$
2. To ensure $f(x)$ is differentiable at $x = u$.
$$\gamma = -\frac{\ln c}{Bc^u},$$

the probability density function becomes

$$f(x) = \begin{cases} \exp\left(-\frac{B}{\ln c}(c^x - 1)\right) Bc^x & \text{if } x \leq u \\ f(u)\left(1 - (x-u)\ln c\right)^{\frac{Bc^u}{\ln c} - 1} & \text{if } x > u \end{cases}. \quad (4)$$

Properties of the proposed threshold life table

- Five parameters are reduced to three free parameters B, c, and u;
- The shape and scale of the GDP distribution is implicitly determined by the threshold and the force of mortality at the threshold;
- The second constraint indicates $\gamma < 0$, which means the GPD in Equation (3) has a right end point $-\sigma/\gamma$
- The limiting age of human beings after imposing the two constraints is $u - \sigma/\gamma = u + \frac{1}{\ln c}$.

The likelihood function

- Use the US mortality data in year 2015 at ages 65 to 105: d_x and E_x, for $x = 65, 66, 67, \ldots, 105$.
- Assume the number of deaths at age x follows a binomial distribution.
- The likelihood function of a threshold life table fitting to the period mortality data is

$$L(B, c, \gamma, \sigma, u) = \prod_{x=65}^{105} \binom{E_x}{d_x} q_x^{d_x} (1 - q_x)^{E_x - d_x} \qquad (5)$$

where $q_x = \Pr(X < x + 1 | X > x) = 1 - S(x+1)/S(x)$ is the probabilities of dying between ages x and $x + 1$

Model 1: two constraint

The likelihood function becomes

$$
\begin{aligned}
& L(B,c,u) \\
= & \prod_{x=65}^{x=m-1} \binom{E_x}{d_x} \left(1 - e^{-Bc^x \frac{c-1}{\ln c}}\right)^{d_x} e^{-Bc^x \frac{c-1}{\ln c}(E_x - d_x)} \\
& \times \binom{E_m}{d_m} \left(1 - e^{-B\frac{c^u - c^m}{\ln c}} \left(1 - (m+1-u)\ln c\right)^{\frac{Bc^u}{\ln c}}\right)^{d_m} \\
& \quad \cdot \left(e^{-B\frac{c^u - c^m}{\ln c}} \left(1 - (m+1-u)\ln c\right)^{\frac{Bc^u}{\ln c}}\right)^{E_m - d_m} \\
& \times \prod_{x=m+1}^{105} \binom{E_x}{d_x} \left(1 - \left(1 - \frac{\ln c}{1-(x-u)\ln c}\right)^{\frac{Bc^u}{\ln c}}\right)^{d_x} \left(1 - \frac{\ln c}{1-(x-u)\ln c}\right)^{\frac{Bc^u}{\ln c}(E_x - d_x)},
\end{aligned}
\tag{6}
$$

where $\lfloor \cdot \rfloor$ denotes the greatest integer function and $m = \lfloor u \rfloor$. For any search number m, the boundaries are $m \leq u < m+1$ and $c \leq e^{\frac{1}{106-m}}$ respectively.

Model 2: no constraint

The likelihood function can be written as

$$\begin{aligned} & L(B,c,u,\gamma,\sigma) \\ = & \prod_{x=65}^{m-1} \binom{E_x}{d_x} \left(1 - e^{-Bc^x \frac{c-1}{\ln c}}\right)^{d_x} e^{-Bc^x \frac{c-1}{\ln c}(E_x-d_x)} \\ & \times \binom{E_m}{d_m} \left(1 - e^{-\frac{B}{\ln c}(c^u-c^x)}(1+\frac{\gamma}{\sigma}(m+1-u))^{\frac{-1}{\gamma}}\right)^{d_m} \\ & \qquad \cdot \left(e^{-\frac{B}{\ln c}(c^u-c^x)}(1+\frac{\gamma}{\sigma}(m+1-u))^{\frac{-1}{\gamma}}\right)^{E_m-d_m} \\ & \times \prod_{x=m+1}^{105} \binom{E_x}{d_x} \left(1 - \left(1+\frac{\gamma}{\sigma+\gamma(x-u)}\right)^{\frac{-1}{\gamma}}\right)^{d_x} \left(1+\frac{\gamma}{\sigma+\gamma(x-u)}\right)^{\frac{-1}{\gamma}(E_x-d_x)}, \end{aligned} \tag{7}$$

where $m = \lfloor u \rfloor$, and $\gamma(106-u)+\sigma > 0$ if γ is negative

Model comparison

To compare the results of fitting threshold life tables to the mortality data, we consider the following three models:

- Model 0: the Gompertz law
- Model 1: the threshold life table with two constraints
- Model 2: the threshold life table without constraints

Maximum likelihood estimates

	Males			**Females**		
	Model 0	Model 1	Model 2	Model 0	Model 1	Model 2
B	1.85E-05	1.91E-05	2.27E-05	5.28E-06	5.55E-06	5.49E-06
c	1.1062	1.1058	1.1033	1.119	1.1183	1.1184
u		100.003	93		103	95
γ			-0.256			-0.252
σ			3.917			3.77
ω		109.95	108.28		111.94	109.96
Logl	-1745.34	-1736.96	-1350.97	-1702.99	-1691.42	-1547.67
BIC	3498.11	3481.35	2709.36	3413.42	3390.27	3102.76

Table: Parameter estimates of the three mortality curves

Likelihood ratio test

	Null	Alternative	χ^2 statistics	p-value
Males	Model 0	Model 1	16.76	4.2E-05
	Model 1	Model 2	771.99	0
Females	Model 0	Model 1	23.15	1.5E-06
	Model 1	Model 2	287.50	0

Table: Likelihood ratio tests for model selection

Graphic comparison

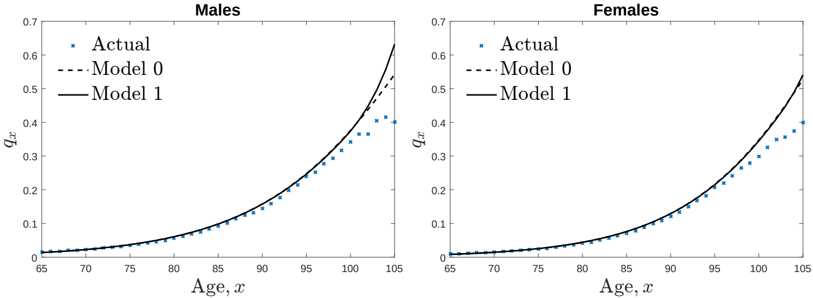

Figure: Raw mortality probabilities, the fitted Gompertz law and the fitted threshold life table with constraints, for males (left) and females (right)

Graphic comparison

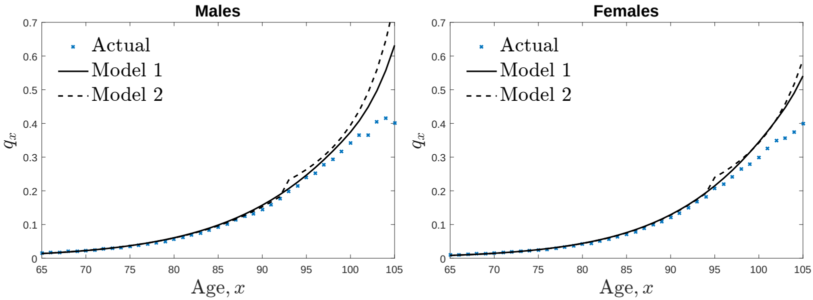

Figure: Raw mortality probabilities and the with and fitted threshold life table without constraints, for males (left) and females (right)

Prior distributions

- Assume the parameters B, c, and u of the threshold life table with smoothness constraints are independent;
- Three independent prior distributions are

$$
\begin{aligned}
f(B) &= \frac{1}{\lambda} e^{-\frac{B}{\lambda}}, \quad B > 0, \lambda > 0, \\
f(c) &= \frac{\theta}{(1 - c_m^{-\theta}) c^{\theta+1}}, \quad 1 < c \leq c_m, \theta > 0, \\
f(u) &\propto (u - m)^{\alpha-1} (m + 1 - u)^{\beta-1}, \quad m \leq u < m + 1, \alpha > 0, \beta > 0,
\end{aligned}
$$

where λ, θ, α, and β are hyper parameters.

Joint posterior probability

- Let **E** and **d** be the vector with elements E_x and d_x respectively, for $x = 65, 66, \ldots, m-1, m, m+1, \ldots, 105$;
- Assuming the counts of death at different ages are independent random variables;
- The joint posterior probability density function of the parameters B, c, u can be written as

$$\begin{aligned} p(B, c, u|\mathbf{d}) &\propto f(B)f(c)f(u)L(B, c, u) \\ &\propto \frac{\theta(u-m)^{\alpha-1}(m+1-u)^{\beta-1}}{\lambda c^{\theta+1}} e^{-\lambda/B} L(B, c, u), \end{aligned}$$

where $L(B, c, u)$ the likelihood function in Equations (6)

Conditional posteriors

The conditional posteriors can be written as

$$f_1(B|c,u,\mathbf{E},\mathbf{d}) \propto e^{-\frac{B}{\lambda}} f(B,c,u,\mathbf{E},\mathbf{d}), \quad (8)$$

$$f_2(c|B,u,\mathbf{E},\mathbf{d}) \propto \frac{1}{c^{\theta+1}} f(B,c,u,\mathbf{E},\mathbf{d}), \quad (9)$$

$$f_3(u|B,c,\mathbf{E},\mathbf{d}) \propto (u-m)^{\alpha-1}(m+1-u)^{\beta-1} f(B,c,u,\mathbf{E},\mathbf{d}), \quad (10)$$

where

$$f(B,c,u,\mathbf{E},\mathbf{d}) = \prod_{x=65}^{x=m-1} \left(1 - e^{\frac{-Bc^x}{\ln c}(c-1)}\right)^{d_x} e^{\frac{-Bc^x}{\ln c}(c-1)(E_x-d_x)}$$
$$\times \left(1 - e^{\frac{-B}{\ln c}(c^u-c^m)}(1-(m+1-u)\ln c)^{\frac{Bc^u}{\ln c}}\right)^{d_m} \left(e^{\frac{-B}{\ln c}(c^u-c^m)}(1 - (m+1-u)\ln c)^{\frac{Bc^u}{\ln c}}\right)^{E_m-d_m}$$
$$\times \prod_{x=m+1}^{105} \left(1 - \left(1 - \frac{\ln c}{1-(x-u)\ln c}\right)^{\frac{Bc^u}{\ln c}}\right)^{d_x} \left(1 - \frac{\ln c}{1-(x-u)\ln c}\right)^{\frac{Bc^u}{\ln c}(E_x-d_x)}$$

Compare MLE and Bayesian estimates

- Following Aminzadeh (2013), we use the Markov chain Monte Carlo (MCMC) method based on the Metropolis & Hastings algorithm to generate samples;
- At each iteration of which all parameters are updated simultaneously;
- Propose to use the following proposal distributions for the parameters B, c, and u.
    - The proposal distribution for B is $EXP(B_{mle})$ an exponential distribution with mean equal to mle of B.
    - The proposal distribution for c is $N(c_{mle}, 0.005^3)$ a normal distribution with mean equal to mle of c.
    - The proposal distribution for u is $N(u_{mle}, 0.1^2)$ a normal distribution with mean equal to mle of u.

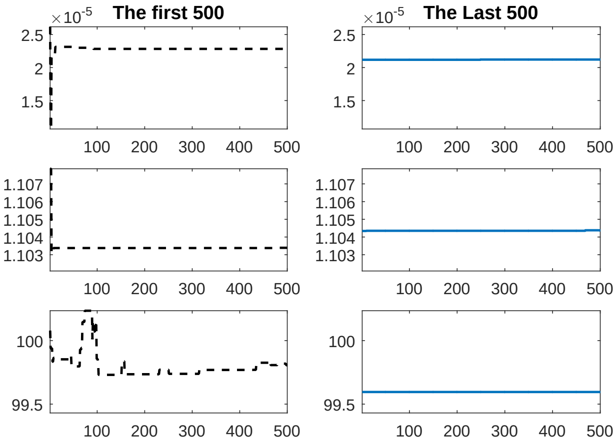

Figure: Markov chain Monte Carlo samples

Simulation results

	True value	MCMC		MLE	
		Mean	MSE	Mean	MSE
B	1.915E-05	1.906E-05	9.53E-12	1.891E-05	4.76E-14
c	1.1058	1.1060	4.648E-06	1.1059	2.45E-08
u	100.003	99.91	0.104	100.218	0.563

Table: Compare the MCMC method and maximum likelihood estimation based on 100 sets of simulated numbers of death data

Conclusions

- Threshold life table with two constraints is favorable model
- The MCMC method performs better than the maximum likelihood estimation
- Predictive analysis of the threshold life tables enables Bayesian inference of the life time distribution, limiting age, and other quantities of interest

Selected references:

1. Aminzadeh, M. S. (2013) "Bayesian estimation of the expected time of first arrival past a truncated time T: the case of NHPP with power law intensity." *Computational Statistics* 28, no. 6 : 2465-2477.
2. Carreau, J., and Bengio, Y. (2009). A hybrid Pareto model for asymmetric fat-tailed data: the univariate case. *Extremes*, 12(1), 53-76.
3. Carriere, J.F. (1994). An investigation of the Gompertz law of mortality. *Actuarial Research Clearing House*, Vol 2.
4. McNeil, A. J., Frey. R., and Embrechts, P. (2015). *Quantitative risk management: Concepts, techniques and tools. Princeton university press.*
5. *Lee, R. D. and Carter, L. (1992). Modeling and forecasting the time series of US mortality. Journal of the American Statistical Association, 87: 659-671.*
6. *Li, J. S.-H., Hardy, M.R. and Tan, K. S. (2008). Threshold life tables and their applications. North American Actuarial Journal,12, 99-115.*
7. *Watts, K. A., Dupuis, D. J. and Jones, B. L. (2006). An Extreme Value Analysis of Advanced Age Mortality Data. North American Actuarial Journal, 10(4), 162-78.*