Liet. matem. rink., T. 40, spec. nr., 486–490

Fitting parametrical models to Lithuanian cause – specific mortality data

Rasa SENKUVIENĖ (LSIC, MII)
e-mail: rasa.miseikyte@takas.lt

1. Introduction

There have been many attempts in recent years to describe the population mortality experience in each age interval by the means of small number of parameters. Best way to do that is to find an appropriate parametrical model, fitting empirical survival function. During last century a big variety of such models were developed. Most of these models were designed to fit the mortality in older ages, so reasonably can be treated as models of senescent mortality. Aim of this study was to test some families of parametric survival functions on Lithuanian 1988 – 1996 years mortality data. Both overall and cause specific partial survival functions were fitted in order to find best approximations for all causes of death mortality (ALL), mortality from cardiovascular diseases (CVD), mortality from cancer(CAN) and for mortality from external causes of death (EXT). Five families of survival functions were analysed, which, of couse, doesn't embrace all great variety of models, developed in recent years. Nevertheless, this work should be viewed as one more effort in the framework of complex analysis of survival functions.

2. Models

Population age specific mortality can be described by the survival function

$$S(t) = \exp\left[-\int_0^t \mu(\tau)\,\mathrm{d}\tau\right], \tag{1}$$

there $\mu(t)$ denotes the *force of mortality* at age t, or, in terms of reliability theory, the *hazard* or *intensity* rate. Cause specific partial survival functions are defined as follows:

$$S_c(t) = \exp\left[-\int_0^t \mu_c(\tau)\,\mathrm{d}\tau\right],$$

there $\mu_c(t)$ denotes the cause specific force of mortality. Consider five families of survival funtions:

$$S_1(t) = \exp\left[-\Lambda(t)\right], \quad S_2(t) = \left[1+\Lambda(t)\right]^{-1}, \quad S_3(t) = \exp\left[1-\exp\left(\Lambda(t)\right)\right],$$

$$S_4(t) = \left[1+\rho^2\Lambda(t)\right]^{-\frac{1}{\rho^2}}, \quad S_5(t) = \exp\left[-\frac{1-\theta}{\theta\rho^2}\left(\left(1+\frac{\rho^2}{1-\theta}\Lambda(t)\right)^{\theta}-1\right)\right].$$

First three families, named *Weibull*, *Logit* and *Bajen* are described in Slimen [1]. Last two ones are mixtures of Weibull family – *Gamma* mixture and mixture, described in Hougaard [2]. There $\Lambda(t)$ is generated by one of nine transformations:

$$\Lambda_1(t) = \exp\left[\alpha+\beta\log(t)\right], \quad \Lambda_2(t) = \Lambda_1\left[\frac{t}{H-t}\right],$$

$$\Lambda_3(t) = \exp\left[\alpha+\beta\frac{t^k-t^{-k}}{2k}\right], \Lambda_4(t) = \exp\left[\alpha+\beta\log\left(\log(1+t^k)\right)\right],$$

$$\Lambda_5(t) = \exp\left[\alpha+\beta\log(e^{kt}-1)\right], \quad \Lambda_6(t) = \Lambda_4\left[\frac{t}{H-t}\right],$$

$$\Lambda_7(t) = \Lambda_3\left[\frac{t}{H-t}\right], \quad \Lambda_8(t) = \Lambda_5\left[\frac{t}{H-t}\right], \quad \Lambda_9(t) = -Q\log\left[1-\frac{e^{\gamma t}-1}{e^{\gamma H}-1}\right],$$

where $0 < \theta < 1, \rho^2 > 0, \alpha, \beta > 0, H > t, k > 0, Q > 0$ and $\gamma > 0$ are parameters and t denotes the age. Let

$$S_{mk}(t) = S_m\left(\Lambda_k(t)\right) \tag{2}$$

to denote the model from m-th family, generated by k-th transformation. Survival functions, generated by (2) formula give a good fits in older ages. They also are useful when modeling cause specific partial survival functions from the causes, prevailing in senility and having small numbers of death in early childhood (CVD, CAN). When fitting all causes survival function two additional parameters were added to fit mortality in early childhood [5].

$$S(t) = (1+\sigma^2\lambda t)^{-\frac{1}{\sigma^2}} S_{mk}(t), \quad \sigma^2 > 0, \quad \lambda > 0.$$

3. Estimation

Distribution of numbers of death m_j in age group $[t_j, t_{j+1})$ can be approximated by a Poisson random value with parameter $n_j q_j$, where n_j denotes number of living at the beginning of age interval $[t_j, t_{j+1})$ and q_j – probability to die, determined by model. So it is meaningfull to evaluate parameter values by minimising likelihood ratio logarithm:

$$\log(L) = 2\sum_{j=1}^{N} m_j \log\left(\frac{m_j}{n_j q_j}\right) - m_j + n_j q_j, \tag{3}$$

where N denotes number of age groups. When model is true, population estimates n_j correct and $N \to \infty$,

$$\log(L) \sim \chi^2(N-n),$$

where n denotes the number of parameters evaluated.

In order to asses overall quality of approximation (joint over 88 – 96 years) one more optimization function was used. Substantiation of this function lays upon assumption that values of empirical probability to die in j-th age group at i-th year $\hat{q}_{ji}$ are distributed nearly identical over years (since the time period from 1988 to 1996 year is too short for big changes in population numbers or in mortality). If $\hat{q}_{ji}$ distributes independently with the same mean and population numbers in each age group remains unchanged over years and equals to n_j, then distribution of $\hat{q}_{ji}$ is asymptotically normal. Then all these assumptions are hold,

$$\rho_j = \frac{9(\hat{q}_{j\cdot} - q_j)^2}{\frac{1}{8}\sum_{i=88}^{96}(\hat{q}_{ji} - \hat{q}_{j\cdot})^2} \sim F(1,8), \tag{4}$$

when $n_j \to \infty$. There $F(1,8)$ is a Fisher's distribution with 1 and 8 degrees of freedom, $\hat{q}_{j\cdot} = \frac{1}{9}\sum_{i=88}^{96}\hat{q}_{ji}$ – mean empirical probability to die in j-th age interval and q_j - probability to die in the same age interval, given by model. Mean of $F(1,8)$ equals to $\frac{8}{6}$. So, when optimization functional is defined as:

$$Q_n = \frac{6}{8\cdot(N-n)}\sum_{j=1}^{N}\rho_j, \tag{5}$$

it should be expected tend toward 1 when numbers of deaths in each age group increases.

In practice numbers of death are too small to give satisfactory approximation of the ρ_j by $F(1,8)$. Moreover, population numbers are influenced by measurement errors and influence of other factors are ignored (reason why (3) optimisation function is not very usefull when errors arrising from other sources in some age groups are much bigger than errors from Poisson distribution). In order to obtain an approximate insight of how values of Q_n is distributed and which values of it can be treated as showing satisfactory approximation, some simulation procedure was used. New samples $(\hat{q}^*_{j1},\ldots,\hat{q}^*_{j9})$ were drawn randomly (with returning) from the actual sample $(\hat{q}_{j88},\ldots,\hat{q}_{j96})$ and then the empirical 5% upper bound of function's

$$Q'_n = \frac{6}{(N-n)}\sum_{j=1}^{N}\frac{9(\hat{q}^*_{j\cdot} - \hat{q}_j)^2}{\sum_{i=1}^{9}(\hat{q}^*_{ji} - \hat{q}^*_{j\cdot})^2}, \tag{6}$$

distribution was calculated. Such simplification of ordinary bootstrap procedure was performed, since "bootstrapping" the Q_n for each nonlinear model requires too much computer time resourses.

4. Results

The results of approximation of the Lithuanian total 88-96 years mortality by function, defined in (2) are presented in tables:

Q_n values for males	ALL	CAN	CVD	EXT
Λ_1	18.29	6.61	5.59	8.39
Λ_2	2.60	6.37	1.21	8.42
Λ_3	2.26	4.27	1.01	6.82
Λ_4	19.26	6.61	5.65	8.55
Λ_5	2.58	4.93	1.25	6.78
Λ_6	2.34	3.90	1.06	8.20
Λ_7	2.62	6.44	1.23	8.54
Λ_8	2.44	4.77	1.05	6.85
Λ_9	2.65	4.97	1.23	7.45
Q'_n 5% upper bound	2.09	4.65	3.18	2.96

Q_n values for females	ALL	CAN	CVD	EXT
Λ_1	27.95	4.42	12.52	6.27
Λ_2	2.77	4.23	2.07	5.07
Λ_3	2.58	2.93	2.19	3.34
Λ_4	28.28	4.46	12.66	6.05
Λ_5	9.57	2.95	2.24	2.90
Λ_6	2.02	2.93	2.10	3.23
Λ_7	2.77	4.27	2.08	5.12
Λ_8	2.60	2.94	2.09	2.78
Λ_9	5.13	3.20	2.29	3.55
Q'_n 5% upper bound	2.16	3.20	2.28	1.90

Note that although fifth family of survival functions, having largest number of parameters and including first, second and forth families as partial cases gives best approximations in most cases, attempts to fit models with less number of parameters is meaningfull in order to avoid hyperparametrisation. Fo example, best fit of cancer mortality for both males and females is given not by S_{56}, but by S_{46} model (values 3.90, 2.93 and 3.86,2.89, respectively).

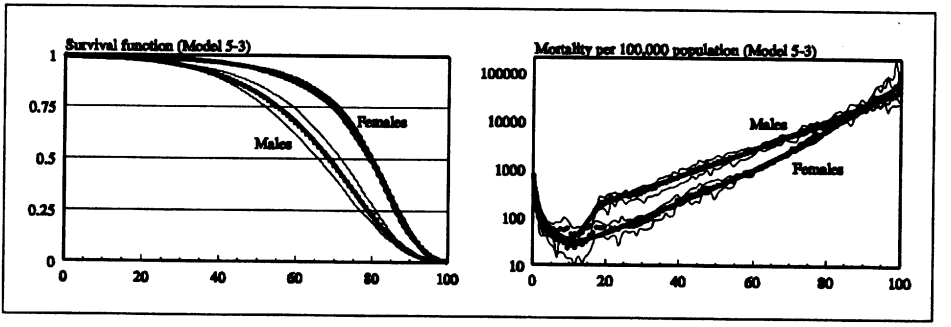

Fig. 1. Minimum, maximum and average empirical values for Lithuanian 1988–1996 year's survival and mortality functions and predictions from model $S(t) = (1 + \lambda\sigma^2 t)^{-\sigma^{-2}} S_{53}(t)$, fitted to the same data. For males(females): $\sigma^2 = 376(475)$, $\lambda = 0.042(0.003)$, $\theta = 0.08(0.29)$, $\rho^2 = 341(85)$, $\alpha = -13.3(-5.2)$, $\beta = 0.585(0.003)$, $k = 1.25(2.18)$

References

[1] D.J. Slimen, P.A. Lasenbruch, Survival distributions arising from two families and generated by transformations, *Commun. Statist.: Theory and Meth.*, **13**(10), 1179–1201 (1984).

[2] P. Hougaard, Fraitly models for survival data, *Lifetime Data Analysis*, **1**, 255–273 (1995).

[3] V.K. Koltover, Z.S. Andrianova, A.N. Ivanova, Modeling of survival and mortality curves of human population based on the theory of reliability, *Nauka, ser. Biol.*, **1** 121–129 (1993) (in Russian).

[4] D. Zelterman, A Statistical distribution with an unbounded hazard function and its application to a theory from demography, *Biometrics*, **48**, 807–818 (1992).

[5] R. Mišeikytė, Lietuvos gyventojų kohortinio išgyvenamumo modeliavimas, *LMD mokslo darbai*, 366–371 (1998).

Lietuvos gyventojų mirtingumo pagal priežastis aproksimavimas parametriniais modeliais

R. Senkuvienė

Lietuvos gyventojų 1988 – 1996 m. mirtingumas pagal pagrindines ligų klases (kraujo apytakos sistemos ligos, piktybiniai navikai bei išorinės mirties priežastys), o taip pat bendras mirtingumas buvo aproksimuoti parametriniais išgyvenamumo funkcijų modeliais. Modelių palyginimui bei aproksimacijos kokybės įvertinimui pateiktos lentelės su kokybės funkcionalo minimumo reikšmėmis.