Using quantile regression to identify longevity thresholds

Nicola Tedesco and Luisa Salaris

Department of Social Sciences and Institutions
University of Cagliari (Italy)

1. Introduction

From the review of studies on longevity emerges that there is not a common agreement on age-threshold for the identification of long-lived populations and individuals. More frequently researchers adopt their strategies, perspective and longevity thresholds according to the available data. However, among the variegate survival cut-off used it is possible to classify longevity thresholds into two large groups, namely “fixed" and "relative" threshold.
The fixed threshold are identified in correspondence of specific ages, for example at 50, 60, 70, 80 years and the choice is entirely arbitrary and heavily depends on the specific research question. For example, researchers interested in investigating the effect of genetic endowment on survival at advanced ages are generally oriented in the study of individuals 50 years and over (Christensen et al. 2006). Therefore the use of a survival threshold is instrumental to the identification of the population understudy.
There are also numerous studies that chose a specific longevity threshold to identify inside a specific population subgroups of individuals, that accordingly to the selected cutoff age are classified as long-living or not.
The applications of “relative” thresholds are the same of the fixed one just discussed, but what changes is the procedure according to which the cutoff age is chosen. In this case the identification of the longevity threshold occurs according to the distribution of deaths and its cumulative percentages - for example, the 8th or the 9th decile (Blackburn et al. 2004).
Both longevity thresholds prove to have strengths and weaknesses and in general, researchers are more oriented to adopted the cutoff age used in similar studies in order to be able to compare results. However, among the possible points of reflection there are two aspects that deserve special attention: i) survival experience of a population along the entire life cycle can greatly differ from another, despite for example reaching similar level of survival at older ages; ii) given the heterogeneous composition of the population, when analyzing differential mortality it could be useful to think in terms of population selection, devoting attention not exclusively to the robust component, the long-living ones, but also to the frail individuals, who exit from the population at earlier stages.
The questions that arise are numerous: why some people died earlier than others? Which variables are involved in the selection process? And the latter have a constant effect on individuals’ survival or their influence varies at different stages of life? And in other words, do the estimated effect of the selected variables vary accordingly to the longevity threshold chosen for investigation?
In the attempt to answer these questions, the present paper proposes the use of Quantile Regression Models (QRM) as a useful method for the identification of longevity threshold as they allow to examine the evolution of survival experience of the population under study at different ages and in the meantime to check the effect of covariates. The use of QRM is applied in the present contribution to the study of Villagrande Strisaili (Italy) population.

2. Data and methods

The database used comes from the *Villagrande Longevity Database* (VILD), which includes all individuals born in Villagrande Strisaili (Italy) from 1866 to 1915. For each individual the exact date at death or the proof that he/she was still alive at the date of investigation have been traced. The data was gathered from civil registers (which record all births, marriages, and deaths), parish registers and the population registers (*anagrafe*). For further details on VILD data see Salaris 2010.

Survival data is here analyzed by means of Quantile Regression Models (QRM). These models could be particularly useful in the study of longevity as they enable to estimate quantiles of age at death as a function of a set of predictors. In this way, according population characteristic and/or structure, we can obtain more precise estimates of parameters. In this work we estimate some important percentiles: 10-th, 20th, 25-th, 50-th, 75-th, 80-th and 90-th to study both usually most import points of deaths distribution and its tails.

QRM are robust models, less sensitive to the presence of outliers and they do not require particular assumptions about the distribution of survival times (Koenker and Bassett 1978). The simultaneous estimation of a set of quantiles using QRM allows to study the effect of predictors for different levels of longevity and to estimate the variance-covariance matrix simultaneously for the different quantiles to obtain optimal confidence intervals of estimated parameters. QRM could also be used to predict every quantile of age of death as a function of a set of values of significance predictors.

QRM were used in different research areas, particularly in ecology and biology (Koenker and Geling 2001; Knight and Ackerly 2002), and most recently in the analysis of the importance of inequality as predictor of mortality rates (Yuang et al., 2012). Algebraically, we have that given a vector $\boldsymbol{Z}$ of covariates and $\boldsymbol{\tau} \in [\mathbf{0}, \mathbf{1}]$, the equation of a QRM that link linaearly $\boldsymbol{Q_Y(\tau|Z)}$ to $\boldsymbol{Z}$ is

$$\boldsymbol{Q_Y(\tau|Z) = Z^T \beta(\tau)}$$

where $\boldsymbol{Q_Y(\tau|Z)}$ is the conditional quantile to covariates and $\boldsymbol{\beta(\tau)}$ is the vector of unknown parameter of the model, that represents the effect of covariates on the $\boldsymbol{\tau}$ quantile of dependent variable and changes according to different quantiles. In presence of incomplete or truncated data as in survival times analysis, there are two approaches: the first proposed by Portnoy (2003) to estimate conditional quantile functions generalizing the Kaplan-Meier method; the second proposed by Peng and Huang (2008) based on Nelson-Aalen estimator of the cumulative hazard function. Indicating with T the survival time, respectively we have

$$\boldsymbol{Q_Y(\tau|Z) = H_0^{-1}\big(-log(1-\tau)\big)e^{-Z\beta}} \text{ where } \boldsymbol{H_0(t) = \int_0^t h_0(s)ds},$$

and

$$\boldsymbol{Q_T(\tau|Z) = exp\{Z^T \beta(\tau)\}}$$

3. Preliminary results

In this paper we applied a QRM for Survival data considering quantile function as a generalization of Kaplan-Meier method. For sake of space we show in Table 1 only models for 10-th, 20-th, 50-th, 80-th and 90-th percentiles. Parameter estimates show that, among the covariates here considered, the age of death of mother is always significant for the estimation of each quantile. In particular, the magnitude of this effect increases up to

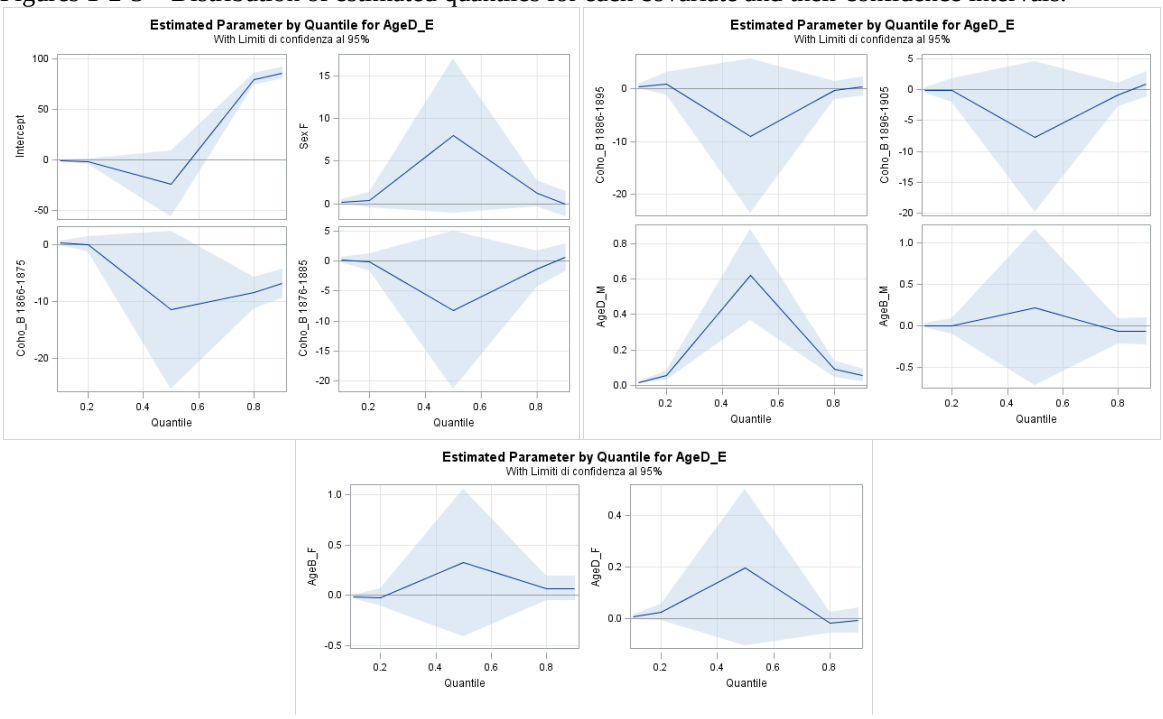
the 50-th percentile, then it decreases. Substantially when the age of death of mother increases, the probability of dying early decreases. Sex, age of death of father and ages at birth of mother and father are not ever significant. It is possible observe a cohort effect only in later ages (80-th and 90-th percentiles), when comparing the first and the last cohort.

Table 1 – Survival QRM for 10-th, 20-th, 50-th, 80-th and 90-th percentiles.

Quantile	Parameters		DF	Stima	Std Error	95% CI		t-value	Pr > |t|
0.1000	Intercept		1	-0.5849	0.4282	-14.243	0.2544	-1.37	0.1721
	Sex	F	1	0.2234	0.1326	-0.0365	0.4834	1.68	0.0922
0.76 yrs	Sex	M	0	0	0	0	0	.	.
	Coho_B	1866-1875	1	0.2900	0.2094	-0.1204	0.7003	1.38	0.1662
	Coho_B	1876-1885	1	0.1639	0.2415	-0.3095	0.6373	0.68	0.4976
	Coho_B	1886-1895	1	0.3461	0.1944	-0.0349	0.7270	1.78	0.0751
	Coho_B	1896-1905	1	-0.1183	0.1894	-0.4895	0.2530	-0.62	0.5325
	Coho_B	1906-1915	0	0	0	0	0	.	.
	AgeD_M		1	0.0163	0.00306	0.0103	0.0223	5.32	<.0001
	AgeB_M		1	0.00268	0.0135	-0.0237	0.0290	0.20	0.8420
	AgeB_F		1	-0.0127	0.0101	-0.0324	0.00706	-1.26	0.2083
	AgeD_F		1	0.00756	0.00398	-0.00025	0.0154	1.90	0.0578
0.2000	Intercept		1	-2.1353	1.4149	-4.9086	0.6379	-1.51	0.1314
	Sex	F	1	0.4387	0.4424	-0.4283	13.057	0.99	0.3215
2.12 yrs	Sex	M	0	0	0	0	0	.	.
	Coho_B	1866-1875	1	0.0977	0.6712	-1.2179	1.4133	0.15	0.8843
	Coho_B	1876-1885	1	-0.1850	0.7574	-1.6695	1.2995	-0.24	0.8071
	Coho_B	1886-1895	1	0.8810	1.0509	-1.1788	2.9407	0.84	0.4020
	Coho_B	1896-1905	1	-0.0641	0.9710	-1.9672	1.8389	-0.07	0.9473
	Coho_B	1906-1915	0	0	0	0	0	.	.
	AgeD_M		1	0.0584	0.0131	0.0327	0.0840	4.46	<.0001
	AgeB_M		1	-0.00644	0.0450	-0.0946	0.0817	-0.14	0.8861
	AgeB_F		1	-0.0226	0.0429	-0.1066	0.0615	-0.53	0.5986
	AgeD_F		1	0.0273	0.0161	-0.00420	0.0588	1.70	0.0895
0.5000	Intercept		1	-23.9277	16.5680	-56.4003	8.5449	-1.44	0.1488
	Sex	F	1	7.9620	4.6061	-1.0658	16.9898	1.73	0.0840
48.55 yrs	Sex	M	0	0	0	0	0	.	.
	Coho_B	1866-1875	1	-11.4976	7.1029	-25.4189	2.4237	-1.62	0.1056
	Coho_B	1876-1885	1	-8.1911	6.7169	-21.3560	4.9738	-1.22	0.2228
	Coho_B	1886-1895	1	-8.9649	7.4149	-23.4979	5.5681	-1.21	0.2268
	Coho_B	1896-1905	1	-7.6769	6.2107	-19.8496	4.4958	-1.24	0.2166
	Coho_B	1906-1915	0	0	0	0	0	.	.
	AgeD_M		1	0.6227	0.1311	0.3658	0.8796	4.75	<.0001
	AgeB_M		1	0.2216	0.4798	-0.7189	1.1621	0.46	0.6442
	AgeB_F		1	0.3222	0.3729	-0.4086	1.0530	0.86	0.3877
	AgeD_F		1	0.1986	0.1540	-0.1031	0.5004	1.29	0.1971
0.8000	Intercept		1	79.8161	2.8487	74.2327	85.3994	28.02	<.0001
	Sex	F	1	12.263	0.7876	-0.3175	27.700	1.56	0.1196
85.14 yrs	Sex	M	0	0	0	0	0	.	.
	Coho_B	1866-1875	1	-8.4651	1.4277	-11.2633	-5.6670	-5.93	<.0001
	Coho_B	1876-1885	1	-1.2985	1.5484	-4.3332	1.7363	-0.84	0.4018
	Coho_B	1886-1895	1	-0.3307	0.8788	-2.0531	1.3917	-0.38	0.7067
	Coho_B	1896-1905	1	-0.8557	0.9494	-2.7165	1.0051	-0.90	0.3675
	Coho_B	1906-1915	0	0	0	0	0	.	.
	AgeD_M		1	0.0917	0.0221	0.0483	0.1351	4.14	<.0001
	AgeB_M		1	-0.0666	0.0759	-0.2154	0.0822	-0.88	0.3806
	AgeB_F		1	0.0683	0.0626	-0.0543	0.1910	1.09	0.2748
	AgeD_F		1	-0.0151	0.0206	-0.0556	0.0253	-0.73	0.4638
0.9000	Intercept		1	85.9288	3.0777	79.8966	91.9609	27.92	<.0001
	Sex	F	1	-0.0279	0.7539	-15.056	14.498	-0.04	0.9705
89.99 yrs	Sex	M	0	0	0	0	0	.	.
	Coho_B	1866-1875	1	-6.8045	1.3017	-9.3557	-4.2533	-5.23	<.0001
	Coho_B	1876-1885	1	0.5960	1.1400	-1.6384	2.8304	0.52	0.6012
	Coho_B	1886-1895	1	0.3681	0.9254	-1.4457	2.1818	0.40	0.6909
	Coho_B	1896-1905	1	0.8833	1.0637	-1.2014	2.9681	0.83	0.4064
	Coho_B	1906-1915	0	0	0	0	0	.	.
	AgeD_M		1	0.0564	0.0189	0.0193	0.0935	2.98	0.0029
	AgeB_M		1	-0.0666	0.0847	-0.2326	0.0994	-0.79	0.4319
	AgeB_F		1	0.0672	0.0612	-0.0528	0.1873	1.10	0.2725
	AgeD_F		1	-0.00535	0.0247	-0.0537	0.0430	-0.22	0.8283

The Figures 1-2-3 show how only for age of death of mother the distribution of estimated quantiles have a particular stretched triangular-shaped distribution and how confidence intervals are narrow, particularly on tails.

Figures 1-2-3 – Distribution of estimated quantiles for each covariate and their confidence intervals.



References

Blackburn M-È., Bourbeau R., Desjardins B. (2004). Hérédité et longévité au Québec ancien. *Cahiers québécois de démographie*, 33(1), 9-28.

Christensen K., Johnson T.E., and Vaupel J.W. (2006). The quest for genetic determinants of human longevity: challenges and insights. *Nature*, 7, 436-48.

Knight C.A., Ackerly D.D. (2002**).** Variation in nuclear DNA content across environmental gradients: a quantile regression analysis. *Ecology Letters* 5.1, 66-76.

Koenker R., Bassett G.W. (1978). Regression Quantiles. *Econometrica*, 46, 33–50.

Koenker R., Geling O. (2001). Reappraising Medfly Longevity: A Quantile Regression Survival Analysis. *Journal of the American Statistical Association*, 96, 458–468.

Peng L., Huang Y. (2008). Survival Analysis with Quantile Regression Models. *Journal of the American Statistical Association*, 103, 637–649.

Portnoy S. (2003). Censored Regression Quantiles. *Journal of the American Statistical Association*, 98, 1001–1012.

Salaris L. (2010). *Searching for longevity determinants: following survival of newborns in a in-land village of Sardinia (1866-2006)*. SIS Best PhD Theses in Statistics and Applications - Demography, CLEUP, Padua (Italy).

Yang T.C. et al (2012), Using quantile regression to examine the effects of inequality across the mortality distribution in the U.S. counties. *Social Science & Medicine*, 74, 1900-1910.