Unification of accelerated and proportional hazard rate models and application for data of the Hungarian National Cancer Registry

Lídia Rejtő*,**

***Abstract*— In the literature of multifactorial survival analysis, individual survival curves are described eventually with a single parameter deeming all survival curves to be parallel ([1], [3], [6], [9]). Previously described theories are not able to produce all form of survival curves we may meet; in contrast, we propose a novel method which can handle cases, for example, with constant hazard rate and rapidly decreasing ones simultaneously. Using our methods we estimated survival chances of 189,026 tumor cases, recognized between 2001 and 2005 in Hungary and recorded in NCR.**

I. INTRODUCTION AND NOTATIONS

Hazard rate arises naturally in lifetime data analysis. When so called explanatory variables or covariates are given with the survival times the the most commonly used model that takes into account the observed values of explanatory variables is the Cox proportional hazard rate model ([3], [11]). Other models are the accelerated hazard rate model and the closely connected accelerated failure time model ([1], [8] [10], [12]).

The conditional hazard rate $\rho(t \mid \mathbf{x})$ in *Cox regression* is equal to $b(\mathbf{x})r(t)$, where
– t is the survival time,
– $\mathbf{x}$ is the vector of explanatory variables,
– r is an integrable positive function,
– b is a multiplicative positive function.
In the dual version called *accelerated hazard rate* $\rho(t \mid \mathbf{x}) = r(a(\mathbf{x})t)$ where
– a is a multiplicative positive function.
Here we shall use the unification of the Cox and accelerated hazard rate models where

$$\rho(t \mid \mathbf{x}) = b(\mathbf{x})r(a(\mathbf{x})t). \tag{1}$$

Suppose we have a right censored sample

$$\mathcal{D}_{\mathbf{x}} = \{(Y_i, \mathbf{x}_i, \varepsilon_i), \quad i = 1, \dots n, \}$$

where $Y_i = \min(T_i, C_i), \quad i = 1 \dots n$ are iid right censored random variables, T_i-s are the iid survival times and C_i-s are the censoring times.
$\mathbf{x}_i$ is a d-dimensional vector of explanatory variables belonging to Y_i, and ε_i stands for censoring.
Let us denote the coefficients in functions a, b by $\alpha_j, \beta_j,\ j = 1, \dots, d.$ Then

$$a(x_i) = \exp\big(\sum_{j=1}^{d} \alpha_j(x_i(j))\big) \tag{2}$$

$$b(x_i) = \exp\big(\sum_{j=1}^{d} \beta_j(x_i(j))\big), \tag{3}$$

where
– d is the number of explanatory variables,
– $x_i(j)$ is the j-th coordinate of $\mathbf{x}_i$, i.e. the value of the jth explanatory variable in the ith sample.
We suppose that the functions $a,\ b$ are loglinear functions of the explanatory variables.

II. ESTIMATIONS IN THE SEMIPARAMETRIC MODEL

Let us consider the right censored sample $\mathcal{D} = (Y_i, \varepsilon_i),\ i = 1, \dots, n,$ from the semiparametric model with hazard rate function (1), $Y_i = \min(T_i, C_i)$, $\varepsilon_i = I(T_i \le C_i)$, where $I(E)$ denotes the indicator function of the event E. The random variables $T_i,\ C_i$ are independent and C_i is the iid censoring sequence. We suppose that the two sequences of random variables are independent of each other. In the model two additional iid sequences of positive variables $A_i, B_i \quad i = 1, \dots, n$ are given where A_i and B_i are iid sequences and they are independent of (Y_i, ε_i). The conditional survival function of T_i, given A_i, B_i, is

$$P(T_i > t \mid A_i, B_i) = S(tA_i)^{B_i/A_i} \tag{4}$$

where $S(t)$ is an unknown continuous survival function. The maximum likelihood estimation of the baseline survival function S is a nonparametric estimation problem.

The expression "likelihood estimation" needs explanation in the nonparametric case, because the likelihood method is applicable when a dominated class of measures depends on some parameter. In our case, we assume that the observation has a probability measure P_F which depends on the unknown distribution (or survival) function F. The class $\{P_F\}$ has no dominating measure so we need a more general definition of maximum likelihood. We use the following definition of generalized maximum likelihood estimator suggested by Kiefer and Wolfowitz in [7]. Different approaches of this problem are discussed for example in [1], [2], [4], [5].

*Alfréd Rényi Mathematical Institute of the Hungarian Academy of Sciences, Budapest, P.O.Box 127, H-1364, Hungary
** Statistics Program, University of Delaware, Newark, DE, 19716 USA
rejto@udel.edu

ISBN 978-963-311-370-7

DEFINITION: Let $\mathcal{C} = \{P\}$ be a class of probability measures on a measurable space $(\Omega, \mathcal{A})$. $\hat{P}$ is generalized maximum likelihood estimator (GMLE), if for any $P \in \mathcal{C}$

$$\frac{d\,\hat{P}(\mathbf{X})}{d\,(\hat{P}+P)} \geq \frac{d\,P(\mathbf{X})}{d\,(\hat{P}+P)}, \tag{5}$$

where $\mathbf{X}$ denotes the given sample.
Notice that if $\hat{P}(\mathbf{X}) > 0$ and $P(\mathbf{X}) = 0$ for the given sample then $\frac{d\,P(\mathbf{X})}{d\,(\hat{P}+P)} = 0$. So to find a GMLE, or to check (5) for $P \in \mathcal{C}$ it is sufficient to consider P-s with $P(\mathbf{X}) > 0$. These measures are positive on the given sample, that is in case of discrete measures (5) reduces to the inequality

$$\hat{P}(\mathbf{X}) \geq P(\mathbf{X}).$$

Let us introduce the accelerated times $U_i = Y_i A_i$ and let us suppose that the U_i-s are ordered and the sequence of U_i-s is monotone increasing. Thus $U_1 \leq \ldots \leq U_n$. Set

$$N(t) = \sum_{j=1}^{n} B_j\, I(U_j > t) + \sum_{j=1}^{n} B_j\, I(U_j = t, \delta_j = 0).$$

THEOREM 1. Based on the sample $(Y_i, \varepsilon_i, A_i, B_i)$ $i = 1, \ldots, n$ the generalized maximum likelihood estimator of the continuous baseline survival function $S(t)$ is

$$S(t) = \prod_{i:\, U_i \leq t, \varepsilon_i = 1} \left(\frac{N(U_i)}{N(U_i) + B_i} \right)^{1/B_i} \tag{6}$$

for $\min\limits_{i:\,\varepsilon_i=1} U_i \leq t < \max\limits_{i:\,\varepsilon_i=1} U_i$, and $S(t) = 1$ if $t < \min\limits_{i:\,\varepsilon_i=1} U_i$.

The proof of the theorem is similar to the proof given for Proposition 2.2 in [11].

Let us consider the model described in the Introduction. Thus we have a right censored sample $\mathcal{D}_{\mathbf{x}}$ and the sequences $A_i = a(x_i)$, $B_i = b(x_i)$ are given by (2), (3).
Let us denote the set of parameters by $\mathcal{P}$:

$$\mathcal{P} = \{(\alpha_j, \beta_j), \qquad j = 1, \ldots d\},$$

and the set of permitted baseline hazard function by $\mathcal{R}$:

$$\mathcal{R} = r(t), \qquad 0 \leq t < +\infty.$$

The maximum likelihood estimation of the baseline survival function and $\mathcal{R}$ is a semi-parametric problem. Its nonparametric step is a straight forward consequence of Theorem 1 and it is formulated in the following two theorems.

THEOREM 2. For given $\mathcal{D}_{\mathbf{x}}$, $\mathcal{P}$ the maximum likelihood estimate of the baseline survival function S is a discrete distribution having positive mass in times with $\varepsilon_i = 1$. Let us introduce the accelerated times $U_i = a_i Y_i$ and let us suppose that $\mathcal{D}_{\mathbf{x}}$ is already ordered according to monotone increasing U_i-s. Then the conditional probability q_i of surviving the time U_i with $\varepsilon_i = 1$ is determined by the equation

$$q_i^{b_i} = \frac{N_i}{N_i + b_i},$$

where $N_i = \sum_{k=i+1}^{n} b_k$.

THEOREM 3. Let us suppose that r is a step function, being constant r_k in intervals (v_{k-1}, v_k) of length Δ_k For given $\mathcal{D}_{\mathbf{x}}$, $\mathcal{P}$ the maximum likelihood estimate of r_k is

$$r_k = \frac{\nu_k}{\sum_{i:U_i > r_{k-1}} b_i \delta_i / a_i}, \tag{7}$$

where

$$\delta_k = \begin{cases} \Delta_k & \text{if} \quad U_i \geq v_k \\ u_i - v_{k-1} & \text{if} \quad v_{k-1} \leq U_i < v_k \\ 0 & \text{if} \quad U_i < v_{k-1} \end{cases} \tag{8}$$

– ν_k is the number of U_i-s such that $v_{k-1} \leq U_i < v_k$.
(If the denominator is zero than r_k is not defined).

In the semiparametric model the maximum likelihood method was used to estimate the model parameters.

III. THE PARAMETRIC MODEL AND APPLICATION FOR DATA OF THE HUNGARIAN NATIONAL CANCER REGISTRY

In survival analysis parametric families are frequently applied (see i.e. [9]). To analyze survival distribution of Hungarian cancer patients using data from the National Cancer Registry of Hungary (NCR) between January 1, 2001 and December 31, 2005, accounting for the effect of thirteen available explanatory variables we used the Gompertz model at first (see [13]) .

The semiparametric model, discussed in this paper was developed to provide a better understanding of the NCR data. Investigating the nonparametric estimate of $r(t)$ it turned out [14] that in case of our data the function $r(t)$ has the analytical form

$$r(t) = c_1(1 + c_3 t)^{-c_2},$$

where $c_1 = 1.440$, $c_2 = 0.885$, $c_3 = 19.188$.

Using the unified model (1) with a and b given in (2), (3) we estimated the coefficients $(\alpha_j(x_i(j)), \beta_j(x_i(j)))$ for the given thirteen explanatory variables of NCR. As an example, the following two table contain results by site of the disease. The tables contain:
- Frequency of each value of the tabulated variable
- The maximum likelihood estimator of the Cox parameter, β of the model
- The maximum likelihood estimator of the accelerated hazard rate parameter, α of the model
- Lower quartiles of the two years survival probability p_1
- Upper quartiles of the two years survival probability p_3
- Description of the value of the variable.

Table 1. Coefficients of the Unified Model

Table 1.a. Site of Disease – Males

Freq.	β	α	p_1	p_3	**Description**
4871	-2.174	-1.687	0.451	0.638	Oral cavity
4809	-1.890	-2.025	0.305	0.501	Nasopharynx
275	-2.655	-2.311	0.537	0.712	Other pharynx
2203	-0.973	-0.956	0.085	0.310	Oesophagus
4485	-1.120	-0.365	0.142	0.433	Stomach
14419	-2.139	-0.985	0.436	0.684	Colon & rectum
2027	-0.883	-0.627	0.101	0.348	Liver
3149	-1.095	-0.479	0.090	0.419	Pancreas
3775	-2.407	-2.228	0.493	0.653	Larynx
22221	-1.510	-0.748	0.168	0.513	Lung
2293	-3.235	-2.283	0.739	0.863	Melanoma
486	-2.719	-2.111	0.614	0.745	Breast
11192	-3.338	-2.290	0.707	0.828	Prostate
2275	-2.779	-1.476	0.820	0.941	Testis
4219	-2.282	-1.357	0.542	0.758	Kidney
5529	-2.451	-1.461	0.576	0.744	Bladder
3409	-1.141	-0.805	0.235	0.503	Brain, nerv
488	-2.138	-0.644	0.659	0.828	Thyroid
374	-2.757	-1.462	0.733	0.906	Hodgkin lymph.
1672	-1.876	-0.455	0.493	0.710	Non-Hodgkin lymp.
537	-1.974	-0.777	0.345	0.573	Multiple myeloma
1768	-1.790	-0.436	0.300	0.670	Leukaemia

Table 1.b. Site of Disease – Females

Freq.	β	α	p_1	p_3	**Description**
1987	-2.640	-1.335	0.639	0.795	Oral cavity
1083	-2.077	-1.396	0.473	0.658	Nasopharynx
152	-2.904	-2.127	0.638	0.794	Other pharynx
574	-1.021	-0.059	0.245	0.525	Oesophagus
3374	-1.233	-0.077	0.213	0.507	Stomach
12894	-2.237	-0.846	0.474	0.715	Colon & rectum
1620	-1.057	-0.420	0.156	0.466	Liver
2788	-1.250	-0.669	0.077	0.371	Pancreas
660	-2.204	-1.293	0.565	0.710	Larynx
11389	-1.684	-0.504	0.305	0.652	Lung
2674	-3.661	-2.138	0.826	0.919	Melanoma
25544	-3.602	-2.202	0.832	0.903	Breast
4191	-2.627	-1.354	0.688	0.825	Cervix uteri
4262	-2.929	-1.439	0.742	0.844	Corpus uteri
4513	-2.079	-1.024	0.427	0.720	Ovary
3031	-2.450	-1.148	0.610	0.803	Kidney
2420	-2.489	-1.129	0.615	0.783	Bladder
3397	-1.261	-0.501	0.315	0.604	Brain, nerv
1598	-2.391	0.193	0.839	0.923	Thyroid
349	-2.659	-0.672	0.794	0.935	Hodgkin lymphoma
1632	-2.247	-0.544	0.575	0.769	Non-Hodgkin lymp.
780	-2.321	-0.792	0.456	0.647	Multiple myeloma
1638	-1.637	0.027	0.296	0.674	Leukaemia

REFERENCES

[1] P.K. Andersen, O. Borgan, R.D. Gill and N. Keiding, *Statistical models based on counting processes*, New York, Springer-Verlag, 1991.

[2] P.J. Bickel, C.A.J. Klaassen, Y. Ritov and J.A. Wellner, *Efficient and adaptive estimation for semiparametric models*, The Johns Hopkins University Press, Baltimore and London, 1993.

[3] D.R. Cox and D. Oakes, *Analysis of survival data*, Chapman and Hall, New York, NY, 1990.

[4] S. Geman and C.R. Hwang, Nonparametric maximum likelihood estimation by the method of sieves, *The Annals of Statistics*, **10**, 1982, pp 401–414.

[5] U. Grenander, *Abstract inference*, Wiley, New York, NY, 1981.

[6] J.D.Kalbleisch and R.L.Prentice: *The statistical analysis of failure time data*, 2nd ed, John Wiley & Sons, New York, 2002.

[7] J. Kiefer, J. and J. Wolfowitz, Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters, *Annals of Mathematical Statistics*, **27**, 1956, pp. 887–906.

[8] B. Lee and H.J.P. Timmermans, A latent class accelerated hazard model of activity episode durations. *Transportation Research Part B: Methodological* **41**, 2007, pp. 426-447.

[9] A.W.Marshall and I.Olkin, *Life distributions, Structure of nonparametric, semiparametric, and parametric families*, Springer-Verlag, New York, 2007.

[10] B. Nan, J.D. Kalbleisch and M. Yu, Asymptotic theory for the semiparametric accelerated failure time model with missing data, *The Annals of Statistics*, **37**, 2009, pp. 2351-2376.

[11] L.Rejtő and G.Tusnády: On the Cox regression, *Asymptotic methods in probability and statistics*, A volume in honour of Miklós Csörgő, Proceedings Volume of ICAMPS'97, (Ed. B. Szyszkowitz) Elsevier Science B.V. 1998, pp. 621-637.

[12] M. Schmid and T. Hothorn, Flexible boosting of accelerated failure time models, *BMC Bioinformatics*, **9**, 2008, pp. 269-281.

[13] G.Tusnády, I.Gaudi, L. Rejtő, M. Kásler and Z. Szentirmay, Survival chances of Hungarian cancer patients calculated from the National Cancer Registry. (in Hungarian) *Hungarian Oncology*, **52**, 2008, pp. 339-349.

[14] G.Tusnády, I.Gaudi, L. Rejtő, M. Kásler and Z. Szentirmay, Unification of proportional and accelerated hazard rate models for prediction of survival chances of Hungarian cancer patients in the National Cancer Registry, (manuscript, submitted for publication) 2009.