Hazard based models and covariates

Mariana Craiu, Elena Corina Cipu and Laura Pânzar

Abstract

The aim of this paper is to establish parametric and semiparametric estimation for models with and without covariates (or explanatory variables) that are related to lifetime analysis.

Mathematics Subject Classification: 62C10, 76A05.
Key words: reliability, hazard function, proportional hazard, covariates.

1 Introduction

Survival analysis examines and models the time it takes for events to occur and focuses on the distribution of survival times.

A representation of the distribution of survival times is the hazard function, which asseses the instantaneous risk of failure.

The hazard function is affected by its operation time and also by the covariates under which it operates. It can be take as the product of an arbitrary and unspecified baseline failure rate $h_0(t)$ and a positive function which incorporates the effects of the covariates (the multiplicative model).

2 Piecewise - constant hazard function

Let T represent the time to failure. $S_T = P(T > t)$ is the survival function $h_T(t) = \lim_{\delta \to 0} \dfrac{P(t \le T < t + \delta \mid T \ge t)}{\delta}$ is the hazard function.

There are various ways in which one can estimate $h(t)$ -the hazard function.

One method is one in which consider the intervals $I_j = (a_{j-1}, a_j]$, $j = 1, 2, \dots, k$ and d_j the number of failure times in I_j. By n_j we denote the failure times that exceed a_{j-1} (or the items at risk at a_{j-1}). Every interval I_j must contain at least one time of failure and so, the number of intervals depend on the number of failure times. So a very simple estimation of $h(t)$ is: $\hat{h}(t) \mid_{I_j} = \dfrac{\hat{H}(a_j) - \hat{H}(a_{j-1})}{a_j - a_{j-1}}$, $j = 1, 2, \dots, k$, where $\hat{H}(t)$ is the empirical cumulative hazard function $\hat{H}(t) = \sum_{j:\, t_j \le t} \dfrac{d_j}{n_j}$.

Proceedings of The 3-rd International Colloquium "Mathematics in Engineering and Numerical Physics" October 7-9 , 2004, Bucharest, Romania, pp. 120-127.

A such estimate for hazard function is good when $h(t)$ is close to linear over $(a_{j-1}, a_j]$.

The estimates of hazard functions are generally based on smoothing. In [6] the hazard functions $h(t)$ based on a censored sample are of the form:

$$\overline{h}(t) = \frac{1}{R}\sum_{j=1}^{k} w(\frac{t-t_j}{b})\hat{h}(t_j)$$

where $t_1 < t_2 < \ldots < t_k$ are the failure times in the sample and $\hat{h}(t_j) = \frac{d_j}{n_j}$, $b > 0$ is a window parameter and $w(u)$ is a pdf that is zero outside the interval [-1,1]. Such an estimation approach requires a large number of failure times.

In [5] the authors used step functions to obtain non-parametric estimates of the baseline hazard function.

In [4] Rosemberg takes hazard function as a linear combination of cubic B-splines.

In the subsequent development, we consider the distribution of T through its hazard piecewise - constant function determined by a specified set of points. We also compare its performance to the Aalen - Nelson estimator.

Suppose that lifetimes for individuals in some population follow a distribution with probability density function $f(t), t > 0$ and lifetimes $t_1, t_2, \ldots, t_n$ are observed in a random sample of n independent individuals.

If t_{max} is the ending time of the study, we partition the interval $(0, t_{max}]$ into k disjoint intervals $(a_{j-1}, a_j]$ we assume that:

$$h(t) = \sum_{j=1}^{k} \lambda_j \mathbf{1}_j(t), \quad \mathbf{1}_j(t) = \begin{cases} 1, & t \in (a_{j-1}, a_j] \\ 0, & t \notin (a_{j-1}, a_j] \end{cases} \tag{2.2.1}$$

$j \in \{1, 2, \ldots, k\}$ and $0 \leq a_0 < \ldots < a_k = t_{max}$.

In this case the pdf of T is:

$$f_T = h(t)e^{-\int_0^t h(u)\,du},\ t > 0 \quad \text{and} \quad S(t) = e^{-\int_0^t h(u)\,du}. \tag{2.2.2}$$

The likelihood of $\lambda_1, \ldots, \lambda_k$ is :

$$\mathcal{L}(\lambda_1, \ldots, \lambda_k \mid t_1, \ldots, t_n) = \prod_{i=1}^{n} f(t_i). \tag{2.2.3}$$

The maximum likelihood estimates are obtained via: $\dfrac{\partial \mathcal{L}}{\partial \lambda_r}(\lambda_1, \ldots, \lambda_k) = 0$, $r \in \{1, 2, \ldots, k\}$ that has the solution:

$$\sum_{i=1}^{n} \frac{\mathbf{1}_r(t_j)}{\lambda_r} = \sum_{\{i|t_i \in (a_{r-1}, a_r]\}} (t_i - a_{r-1}) + \sum_{\{i|t_i > a_r\}} (a_r - a_{r-1}). \tag{2.2.4}$$

Denoting by e_r the right side of the equation (2.2) the estimate of λ_r is:

$$\hat{\lambda}_r = \frac{\sum_{i=1}^{n} \mathbf{1}_r(t_i)}{e_r}, \quad r \in \{1, 2, \ldots, k\}. \tag{2.2.5}$$

e_r is named "exposure" and gives the sum of all times by each item in this interval.

3 A semiparametric estimation of $h(t \mid \mathbf{x})$ in the case of a covariate $\mathbf{x}$

Let $\mathbf{x}$ be a covariate vector $\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_s)$ and x_{li} the level "i" of the covariate x_l associated with the value of t_i ($l = 1, \dots, s$; $i = 1, \dots, n$).

We work in the same hypothesis of a picewise constant hazard rate.

The sample is made from the population $T:\ t_1, \dots, t_n$ and the vector $\lambda_1, \dots, \lambda_k$ is the parameter in the life distribution. In this case:

$$h(t) = \sum_{j=1}^{k} \lambda_j \exp(\sum_{l=1}^{s} \beta_{jl} x_l) \mathbf{1}_j(t) \tag{3.3.1}$$

where λ_j and β_{jl}, $j = 1, 2, \dots, k$; $l = 1, 2, \dots, s$ are unknown.

The likelihood is:

$$\mathcal{L}(\lambda_1, \dots, \lambda_k, \beta_{j1}, \dots, \beta_{js} \mid t_1, \dots, t_n) = \prod_{i=1}^{n} \left(h(t_i) \exp(-\int_0^{t_i} h(u)\, du) \right)$$

and the estimates are the solution of the system:

$$\frac{\partial \ln \mathcal{L}}{\partial \lambda_j} = 0, \quad \frac{\partial \ln \mathcal{L}}{\partial \beta_{jl}} = 0, \quad j = 1, \dots, k; \;\; l = 1, \dots, s$$

that gives:

$$\left\{ \begin{aligned} \frac{d_j}{\hat{\lambda}_j} &= \sum_{i=1}^{n} \exp(\sum_{l=1}^{s} \hat{\beta}_{jl} x_{li}) \int_0^{t_i} \mathbf{1}_j(u) du \\ \sum_{i=1}^{n} x_{li} \mathbf{1}_j(t_i) &= \sum_{i=1}^{n} \hat{\lambda}_j x_{li} \exp(\sum_{l=1}^{s} \hat{\beta}_{jl} x_{li}) \int_0^{t_i} \mathbf{1}_j(u) du \end{aligned} \right. \tag{3.3.2}$$

$j = 1, \dots, k$; $l = 1, 2, \dots, s$. In the special case of only one covariate this system become:

$$\left\{ \begin{aligned} \hat{\lambda}_j &= \frac{d_j}{\sum_{i=1}^{n} \exp(\beta_j x_i)[(t_i \wedge a_j - a_{j-1}) \vee 0]} \\ \sum_{i=1}^{n} x_i \mathbf{1}_j(t_i) &= \sum_{i=1}^{n} \hat{\lambda}_j x_i \exp(\hat{\beta}_j x_i)[(a_j \wedge t_i - a_{j-1}) \vee 0]\,. \end{aligned} \right. \tag{3.3.3}$$

When $\beta_j = \beta$ for all $j = 1, \dots, k$

$$\begin{aligned} h(t) &= \sum_{j=1}^{k} \tilde{\lambda}_j \exp(\tilde{\beta} x) \mathbf{1}_j(t) = \exp(\tilde{\beta} x) \sum_{j=1}^{k} \tilde{\lambda}_j \mathbf{1}_j(t) \\ &\text{or} \\ h_i(t) &= e^{\tilde{\beta} x_i} \sum_{1}^{k} \tilde{\lambda}_j \mathbf{1}_j(t) = e^{\tilde{\beta} x_i} h_0(t) \quad \text{(Cox model).} \end{aligned} \tag{3.3.4}$$

In this case the system (3.3) become:

$$
\begin{aligned}
\tilde{\lambda}_j &= \frac{d_j}{\sum_{i=1}^{n} \exp(\tilde{\beta} x_i)[(t_i \wedge a_j - a_{j-1}) \vee 0]} \\
&\text{and} \\
\sum_{i=1}^{n} x_i \mathbf{1}_j(t_i) &= \sum_{i=1}^{n} \tilde{\lambda}_j x_i \exp(\tilde{\beta} x_i)[(t_i \wedge a_j - a_{j-1}) \vee 0]\,.
\end{aligned}
\tag{3.3.5}
$$

Further, one can verify the null hypothesis:

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k$

with respect to the alternative

$H_1 : (\exists) j \neq l$ such that $\beta_j \neq \beta_l$

using the likelihood ratio test by which the critical interval is (C, ∞) where C is given by:

$$
\frac{\mathcal{L}(\tilde{\beta}, \tilde{\lambda}_1, \ldots, \tilde{\lambda}_k \mid t_1, \ldots, t_n)}{\mathcal{L}(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k \mid \hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_k)} \geq C,
\tag{3.3.6}
$$

and C is the percentile of $\mathcal{X}^2_{(k-1)}$ S_0,$F_{\mathcal{X}^2_{(k-1)}}(C) = 1 - \alpha$.

<u>Remark</u>: The mean remaining lifetime $m(t)$, if beta-s are equals, is:

$$
m(t) = \frac{\displaystyle\int_t^{\infty} \exp(-\sum_{j=1}^{k} \tilde{\lambda}_j [(u \wedge a_j - a_{j-1}) \vee 0]) du}{\exp(-\sum_{j=1}^{k} \tilde{\lambda}_j [(t \wedge a_j - a_{j-1}) \vee 0])}
\tag{3.3.7}
$$

4 Analisys of Stone's data in a life test for some specimens of solid epoxy electrical insulation

The failure time T (in minutes) at three levels of voltage (in kV) are given in the following table (Table 1.):

t_i	114	132	144	162	168	174	222	234	245	246
x_i	55	55	55	55	57.5	57.5	55	57.5	52.5	52.5
t_i	252	258	288	288	294	300	312	348	350	390
x_i	57.5	55	57.5	57.5	57.5	55	55	57.5	52.5	57.5
t_i	396	408	444	444	498	510	520	528	546	550
x_i	55	57.5	55	57.5	55	57.5	55	57.5	57.5	52.5
t_i	558	600	690	696	714	740	745	745	772	900
x_i	57.5	52.5	57.5	57.5	57.5	52.5	52.5	55	55	57.5
t_i	1000	1010	1190	1225	1240	1266	1390	148	1464	1480
x_i	57.5	52.5	52.5	52.5	55	55	52.5	52.5	55	52.5
t_i	1690	1740	1805	2440	2450	2600	3000	4690	6095	6200
x_i	52.5	55	52.5	55	52.5	55	52.5	52.5	52.5	52.5

The interval should be specified independently of the data but on the other hand from the data description results that each interval must contain at least one time of failure;

a) For the value of the voltage $v_0 = 52.5$ kV some estimation for $h(t)$ are given in the next tables 2, 3, 4 for different choices of intervals.

The estimated error between our method and Nelson - Aalen method (ϵ). The endpoints of intervals are from the set of sample data.

Table 2. 12 intervals with number of items in each interval randomly choosen:

$\epsilon = 4.52E - 05$

j	$(a_{j-1}, a_j]$	e_j	d_j	n_j	λ_j	λ_j Nelson- Aalen
1	(0,245]	4900	1	20	2.04E-04	2.04E-04
2	(245,350]	1891	2	19	1.06E-03	1.00E-03
3	(350,600]	4200	2	17	4.76E-04	4.714E-04
4	(600,745]	2170	2	15	9.22E-04	9.20E-04
5	(745,1190]	5605	2	13	3.57E-04	3.46E-04
6	(1190,1225]	385	1	11	2.60E-03	2.60E-03
7	(1225,1458]	2262	2	10	8.84E-04	8.58E-04
8	(1458,1690]	1646	2	8	1.22E-03	1.08E-03
9	(1690,1805]	690	1	6	1.45E-03	1.45E-03
10	(1805,3000]	5425	2	5	3.69E-04	3.35E-04
11	(3000,4690]	5070	1	3	1.97E-04	1.97E-04
12	(4690,6200]	2915	2	2	6.86E-04	6.62E-04

Table 3. 12 intervals with the same lenght of intervals $l = 516.667$:

$\epsilon = 4.52E - 05$

j	$(a_{j-1}, a_j]$	e_j	d_j	n_j	λ_j	λ_j Nelson- Aalen
1	(0,516.667]	9624.33	3	20	3.12E-04	2.90E-04
2	(516.667,1033.33]	7261.67	5	17	6.89E-04	5.69E-04
3	(1033.33,1550]	5193	5	12	9.63E-04	8.06E-04
4	(1550,2066.67]	2978.33	2	7	6.72E-04	5.53E-04
5	(2066.67,2583.33]	2450	1	5	4.08E-04	3.87E-04
6	(2583.33,3100]	1966.67	1	4	5.08E-04	4.84E-04
7	(3100,3616.67]	1550	0	3	0.00E+00	0.00E+00
8	(3616.67,4133.33]	1550	0	3	0.00E+00	1.08E+00
9	(4133.33,4650]	1550	0	3	0.00E+00	0.00E+00
10	(4650,5166.67]	1073.33	1	3	9.32E-04	6.45E-04
11	(5166.67,5683.33]	1033.33	0	2	0.00E+00	0.00E+00
12	(5683.33,6200]	928.333	2	2	2.15E-03	1.94E-03

Table 4. 12 intervals with number of items in each interval randomly choosen: $\epsilon = 7.45E-05$

j	$(a_{j-1}, a_j]$	e_j	d_j	n_j	λ_j	λ_j Nelson- Aalen
1	(0,298]	5855	2	20	3.42E-04	3.36E-04
2	(298,575]	4736	2	18	4.22E-04	4.01E-04
3	(575,742.5]	2535	2	16	7.89E-04	7.46E-04
4	(742.5,1100]	4560	2	14	4.39E-04	4.00E-04
5	(1100,1207.5]	1272.5	1	12	7.86E-04	7.75E-04
6	(1207.5,1307.5]	1017.5	1	11	9.83E-04	9.09E-04
7	(1307.5,1469]	1525	2	10	1.31E-03	1.24E-03
8	(1469,1585]	823	1	8	1.22E-03	1.08E-03
9	(1585,2127.5]	3037.5	2	7	6.58E-04	5.27E-04
10	(2127.5,3845]	6347.5	2	5	3.15E-04	2.33E-04
11	(3845,5392.5]	3940	1	3	2.54E-04	2.15E-04
12	(5392.5,6200]	1510	2	2	1.32E-03	1.24E-03

b) When the value of the covariate (the voltage) has the levels $v_1 = 55$ kV and $v_2 = 57.5$ kV, $h(t) = \sum_{j=1}^{k} \lambda_j \exp(\beta_j(v - v_0))\mathbf{1}_j(t)$ with $v_0 = 52.5$ kV. The results of our method are given in the table 5.

Table 5. 10 intervals with different lenght of intervals

j	$(a_{j-1}, a_j]$	d_j	β_j	λ_j
1	(0,174]	6	-0,381087718	0.003685502
2	(174,234]	2	-0,157997523	0.00193278
3	(234,288]	4	0,273394108	0.000816916
4	(288,348]	4	-0,428570459	0.014914637
5	(348,408]	3	0,130518763	0.001366649
6	(408,498]	3	-0,413389195	0.008440873
7	(498,546]	4	0,252590144	0.002029796
8	(546,745]	4	0,345630635	0.000423558
9	(745,1000]	3	0,205763792	0.000754866
10	(1000,2600]	5	-0,514430098	0.011191412

When $\beta_1 = \beta_2 = \cdots = \beta_k$,

$$h(t) = e^{\beta(v-v_0)} \sum_{j=1}^{k} \lambda_j \mathbf{1}_j(t) = e^{\beta(v-v_0)} h_0(t)$$

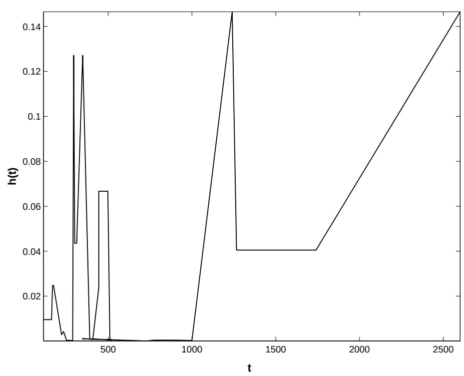
and so, the model is the Proportional hazards model defined by Cox in which covariates have a multiplicative effect on the hazard function.

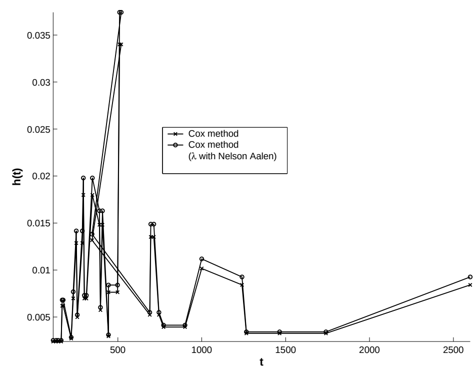
Cox suggested the following likelihood function for estimating β

$$L(\beta) = \prod_{i=1}^{k} e^{\beta x_i} / \sum_{j \in R_i} e^{\beta x_j}$$

where x_i (for us $v_i - v_0$) is the covariate associated with the individual observed to die at t_i and R_i denote the set of individuals who are alive just prior to time t_i.

When β has the same value, for our method and method of Cox the results are: $\beta = -0.37939357757568$, respectively $\beta = -0.39853799343109$. Finaly we present the hazard function in the two cases.





References

[1] Lawless J.F., *Statistical Models and Methods for Lifetime Data*, J. Wiley, 2003.

[2] John D. Kalbfleisch, Ross L. Prentice, *The Statistical Analysis of failure Time Data*, J. Wiley, 2002.

[3] M.J. Crowden, *Statistical Analysis of Reliability Data*,Chapmann Hall, 1991.

[4] Rosenberg P.S. , *Hazard function estimation using B-splines*, Biometrics, 51 (1995), 874-887.

[5] Whitemore A.S., Keller J.B., *Survival estimation using splines*, Biometrics, 42 (1986), 495-506.

[6] Ramlan - Hansen, H., *Smooting counting process intensities by means of kernel functions*, Annals of Statistics, 11 (1983), 453-466.

Mariana Craiu, Elena Corina Cipu and Laura Pânzar
University Politehnica of Bucharest, Department Mathematics III,
Splaiul Independenţei 313, RO-060042 Bucharest, Romania
e-mail: craiu@math.pub.ro, corinac@math.pub.ro, corinac@math.math.unibuc.ro