Statistica Applicata Vol. 8, n. 1, 1996

A HIERARCHICAL MODEL TO IDENTIFY PROGNOSTIC FACTORS IN SURVIVAL ANALYSIS

Paolo Giudici

Dipartimento di Economia Politica e Metodi Quantitativi, Università di Pavia.

SUMMARY

One of the most popular models in the analysis of survival data is Cox's model. Our claim is that, particularly in the preliminary *stage of survival analysis, when explanatory variables need to be selected and properly designed, setting a very complex proportional hazard model may not be the best strategy. We propose an alternative methodology which allows to evaluate, in a simple and explanatory fashion, the relative importance of each potential prognostic factor.*

Keywords: hierarchical partition models, posterior probabilities, survival analysis.

1. BACKGROUND

Survival analysis is often concerned with analysing collections of dependent outcomes. See, for instance, Kalbfleisch and Prentice (1980, hereafter KP), whose terminology and notation will be adopted here.

Consider a collection of survival times, possibly censored, so that the evidence can be represented as $\{y_i = (t_i, \delta_i)\}$, for $i = 1,\ldots,n$, with t_i the observed survival times and δ_i the usual indicator of censoring. Assume that each T_i is distributed according to a c.d.f. $F_i(t)$, with a density $f_i(t)$, from which the (subject–specific) hazard function $\lambda_i(t)$ can be derived as:

$$\lambda_i(t) = \frac{f_i(t)}{1 - F_i(t)}.$$

The main objective of survival analysis is to obtain an estimate of the hazard function, to be subsequently employed, for instance, to predict survival times of patients not included in the study. In order to investigate the possible dependencies among the observations, a typical strategy is to build up a *causal* model which relates the survival times to a proper collection of covariates, say $\underline{Z} = (Z_1,\ldots,Z_k)$,

whose realizations $\underline{z}_i$, for $i = 1,\ldots,n$, are known. The most frequently employed of such models is Cox's proportional hazard model, in which each hazard function is related, conditionally on $\underline{Z} = \underline{z}$, to a common structure:

$$\lambda_i(t / \underline{z}) = \lambda_0(t) exp(\underline{z}\,\underline{\beta}), \tag{1}$$

with $\underline{\beta} = (\beta_0,\ldots,\beta_k)'$ a vector of unknown parameters and $\lambda_0(t)$ a baseline hazard. A typical inference in a Cox model (see e.g. KP) includes selecting a linear combination of $\underline{Z}$ to maximise the likelihood function. The resulting estimate of $\underline{\beta}$ relates the variables in $\underline{Z}$, *assumed* to be prognostic, to the hazard function.

A very crucial aspect of causal models in survival analysis is the preliminary stage, in which a set of explanatory variables must be properly chosen and designed, usually among a very large number of alternatives. This part of the analysis is typically accomplished with the help of *descriptive* tools, such as plots of the observed hazard rates against the covariates values. However, it is often the case that such tools are not sufficiently informative. As a consequence, a large number of variables are included in $\underline{Z}$ and a model selection procedure needs to be run in order to find a parsimonious linear combination.

Our claim is that, in the *exploratory* stage described above, setting a very complex proportional hazard model may not be the best strategy. Some criticisms are:

- When many explanatory variables, possibly correlated, are specified, the *efficiency* of Cox's model selection and estimation becomes heavily dependent on the number of available observations.
- It may be difficult, particularly in observational studies, to have *complete* information on all covariates. Missing data occurs frequently in survival analysis.
- Even when all relevant covariates are included, random effects, expressing accident proneness or *frailties,* induce further dependencies and affect inferences on fixed effects.
- Finally, the aim of model selection is typically to choose a correct model, namely, an appropriate linear combination of the covariates to fit the observed data. Inferences on quantities of interest, such as $\lambda(t|z)$ are then made *conditionally* upon the selected model. Consequently, model uncertainty is not taken into account and, thus, inference may be seriously biased.

The above considerations lead to consider *less structured* methodologies to analyse survival data. In this paper we consider a *collection* of partial exchangeability patterns for the individual hazard functions. Each pattern corresponds to a partition of the individuals, say g, as suggested by the levels of an entertained explanatory

variable. The model we are considering belongs to the class of *hierarchical partition models* defined in Consonni and Veronese (1995). Other important references of our work are Clayton (1991) and Raftery *et al.* (1995).

The plan of the paper is as follows. Section 2 introduces, with reference to an exponential survival time, the basic model. In section 3 we outline the strategy for achieving our two main objectives, namely, evaluation of prognostic factors and estimation of the hazard function. Section 4 illustrates the methodology, with reference to the well–known Veteran's Administration lung cancer data–set (see e.g. KP). Comparisons will be made with exponential regression models, both classical and Bayesian. Finally, section 5 contains some concluding remarks as well as possible extensions.

2. THE PROPOSED MODEL

To illustrate our methodology, we shall assume, in this paper, an hazard function which is constant in time, namely, for $i = 1,...,n$:

$$\lambda_i(t) = \lambda_i . \tag{2}$$

Such a specification, equivalent to assuming an exponential survival model, will enable us to derive simpler results, typically in closed–form and, thus, should permit an easier understanding of the proposed methodology. However, it does not affect the generality of our method, which can be extended to the whole class of parametric survival models.

Having specified the hazard function, it can be shown that, given the evidence $\underline{y} = (y_1,\ldots,y_n)$, the likelihood of $\underline{\lambda} = (\lambda_1,\ldots,\lambda_n)$ is:

$$L(\underline{\lambda}) = \left(\prod_{i \in \mathcal{U}} \lambda_i \right) exp\left\{ -\sum_{i=1}^{n} \lambda_i t_i \right\}, \tag{3}$$

where $\mathcal{U} = \{i: \delta_i = 1\}$ are the uncensored subjects.

Let now g indicate a *partition* of the index set $I = \{1,\ldots,n\}$, with d_g subsets $S_k(g)$, for $k = 1,\ldots,d_g$. Clearly, given the correspondence between I, $\underline{y}$ and $\underline{\lambda}$, g also defines a partition of the data and of the hazard functions. Furthermore, to ease the notation, the dependence of the subsets S_k upon g will be, when possible, dropped.

Notice that the likelihood in (3) assumes all λ_i to be distinct and, thus, is in fact conditional on the *independence* partition $g_{ind} = \{\{1\}, \{2\},\ldots,\{n\}\}$, containing $d_g = n$ separate subsets S_i, each with $n(S_i) = 1$ observations. For this reason, it can be indicated by $L(\underline{\lambda} \mid g_{ind})$.

A different likelihood arises when all hazards can be retained equal to a common rate, say μ. This situation occurs when *no* covariate or frailty affect the

survival times and corresponds to consider all data to be *exchangeable.* The resulting likelihood can be seen as conditional on the partition $g_{exc} = \{1,\ldots,n\}$, containing a single subset S_1 (with $n(S_1) = n$):

$$L(\mu \mid g_{exc}) = \mu^d \exp\{-\mu V\}, \tag{4}$$

where $d = \sum_{i=1}^{n} \delta_i$ represents the total number of failures and $V = \sum_{i=1}^{n} t_i$ the overall time at risk.

Apart from the above situations, which can be regarded as somewhat extreme, survival analysis is typically concerned whith a plurality of effects which may induce dependencies among survival times. Such effects may be either observable (possibly with some missing value) or unobservable. In any case, when relevant, they *induce* a partition of the observations, by associating different hazards to individuals having the same level of the factor.

In our exploratory approach, we shall entertain several partition structures, each induced by the levels of a potential prognostic factor. This amounts to considering a collection of alternative *partial exchangeability structures* for the survival times. Our model consists of two parts: a *likelihood* specification and a *hierarchical prior* distribution on the partition structure as well as on the corresponding set of hazards.

Conditionally on a *general* partition g, let $\lambda_i = \mu_k$, $\forall i \in S_g(g)$. Consequently, the likelihood of the hazards $\underline{\mu} = \left(\mu_1,\ldots,\mu_{d_g}\right)$ is the following:

$$L(\underline{\mu} \mid g) = \prod_{k=1}^{d_g} \mu_k^{d_k} \exp\{-\mu_k V_k\}, \tag{5}$$

where, for $k = 1,\ldots,d_g : \mu_k, d_k = \sum_{i \in S_k(g)} \delta_i$ and $V_k = \sum_{i \in S_k(g)} t_i$ are the hazard, death and risk set of the kth partition subset.

On the other hand, the prior specification requires the definition of a class of possible partitions $\mathcal{G} = \{1,\ldots,G\}$. For instance, in a very preliminary approach, each g can correspond to the observed levels of the entertained effect, either observed or hypothetical. Sebsequently, more elaborate partitions may be considered, such as those resulting from looser categorisations or upon explicit consideration of interaction effects.

Once $\mathcal{G}$ is specified, it is necessary to assign a probability distribution on both $\underline{\lambda} \mid g \in R^{d_g}$ and $g \in \mathcal{G}$. Specifically, conditionally on a partition g we shall take, for $k = 1,\ldots,d_g$ and $\forall i \in S_k(g)$:

$$\mu_k \overset{ind}{\sim} Gamma(r_k m_k, r_k), \tag{6}$$

with m_k and r_k known positive constants. Such constants can be specified using subject–matter knowledge, such as coming from a meta–analysis on the same problem.

The above specification amounts to assume that all individuals in the same partition subset have the same hazard, described by a Gamma random variable with expected value equal to m_k and variance equal to m_k/r_k. Notice also that individuals included in different partition subsets are assumed to have *independent* hazards.

Finally, a simple distribution on $\mathcal{G}$ would take $p(g)$ to be uniformly spread among partitions, i.e. $p(g) = G^{-1}$.

3. COMPUTATIONS

Having specified our exploratory partition model, in this section we outline the relevant computations, in terms of our objectives.

Our first aim is to evaluate the importance of each prognostic factor. This can be achieved calculating, given the observed evidence $\underline{y}$, the posterior probability of each partition, $p(g \mid \underline{y})$.

Following (5) and (6) it can be shown that:

$$p(\underline{y} \mid g) = \prod_{k=1}^{d_g} \frac{(r_k)^{r_k m_k}}{\Gamma(r_k m_k)} \frac{\Gamma(r_k m_k + d_k)}{(V_k + r_k)^{r_k m_k + d_k}}. \tag{7}$$

Finally, application of Bayes' theorem gives $p(g \mid \underline{y}) \propto p(\underline{y} \mid g) p(g)$..

Our second aim is to estimate the hazard function, in order to make predictions on survival times. This task can be performed in two steps: first we work conditionally on a partition, and determine a Bayesian estimate of each individual hazard, by calculating the posterior mean $E(\lambda_i \mid \underline{y}, g)$.

Computationally, following (5) and (6), it turns out that, for $i \in S_k(g)$:

$$E(\lambda_i \mid \underline{y}, g) = \frac{r_k m_k + d_k}{V_k + r_k}. \tag{8}$$

The above expression shows that r_k and $r_k m_k$ can be interpreted, respectively, as pre–experimental "total time at risk" and "observed events". When no prior information is available, they may be taken in an appropriate *uniformative* manner. For instance, in Section 4 we have set all of them equal to 1.

The second step of the estimation procedure involves using $p(g \mid \underline{y})$ to

calculate the marginal posterior expectation of each individual hazard $E(\lambda_i \mid \underline{y})$, via the law of total probabilities:

$$E(\lambda_i / \underline{y}) = \sum_{g=1}^{G} E(\lambda_i / g, \underline{y}) p(g / \underline{y}). \quad (9)$$

As shown, for instance, in Raftery *et al.* (1995), using the marginal posterior expectation via the above model averaging procedure leads to predictions better than those based on conditioning on a single partition, such as that associated to the "best" model.

4. APPLICATION

In this section we apply the proposed exploratory partition model to the well–known Veteran's Administration lung cancer data–set, reported by Prentice (1973).

The above Author and several others following him have concluded that such a data–set is well described by a constant hazard and have accordingly performed model selection and estimation. In describing our results, we shall mainly refer to the classical analysis of Prentice (1973) as well as to the Bayesian analysis of Raftery *et al.* (1995).

Six explanatory covariables were considered as potential prognostic effects for the exponential survival times: performance status, months from diagnosis, age, prior therapy, cell type and treatment. We have first considered the 6 partitions induced by the levels of the available covariates. We have also included those corresponding to complete exchangeability and independence of the survival times. Following (7), and taking $p(g) = 1/8$, we have calculated the posterior probabilities $p(g|\underline{y})$, for $g = 1,\ldots,8$. Such probabilities are reported in Table I.

Tab. I: Posterior probabilities associated to the entertained partitions.

Partition	d_g	$p(g \mid \underline{y})$
g_{exc}	1	.714
g_{ind}	137	10^{-229}
g_{perf}	12	.285
g_{diag}	28	10^{-33}
g_{age}	40	10^{-32}
g_{ther}	2	10^{-12}
g_{cell}	4	10^{-6}
g_{trt}	2	10^{-11}

From Table I it turns out that the highest probability (about .714) is associated to $g_{exc.}$ Concerning the covariables, the results show that, while the partition induced by performance status receives a posterior probability of about .285, all the remaining variables (included treatment) seem *not* to be relevant.

Our conclusions are similar to those in Prentice (1973) and Raftery *et al.* (1995) except for the weaker importance we give to the cell variable. In order to investigate such difference, we further considered different categorisations for the variable *cell.* More precisely, we have considered the 15 partitions which result by considering all the possible contrasts between the cell levels. The obtained results show that the partition with the highest probability corresponds to a contrast between squamos and large cells against small and adeno cells. Its posterior probability results to be about 104 times higher than the original partition and only about 20 times lower than the exchangeability partition. Thus, there appears to be a cell effect on the hazard function, but such an effect requires a *proper classification* of the variable. Notice that our result agrees with what could be obtained by simply looking at the plot of the observed hazard rate d_k/v_k versus the four cell levels.

To summarize, the hazard function can be estimated, following (9), as a mixture of three effects: *a general effect,* corresponding to exchangeability, a *performance effect,* with 12 different estimated values and, finally, a dichotomous *cell effect,* with corresponding posterior probabilities equal to .69, .28, .03. In particular, the estimated hazard contribution of the latter is equal to about .005 for squamos or large cells and to .013 for small or adeno cells.

5. CONCLUDING REMARKS

The exploratory partition model presented here provides a simple and general tool to perform a preliminary evaluation of prognostic effects in survival analysis.

The resulting methodology requires to associate a partition to each tentative explanatory variable. Although this may involve the delicate step of categorising a variable, it has the advantage of being a simple and general procedure. For instance, such a specification may correspond to: a) a fixed covariate, properly stratified; b) a frailty variable, defined in terms of random effects; c) an incomplete covariate, provided that all individuals with missing values are put in the same group.

Possible extensions of the methodology, in the area of survival analysis, include dealing with more complex hazard functions and add further layers to the hierarchical prior. These extensions are generally possible, although Markov chain Monte Carlo methods would be extensively required to perform the relevant computations.

ACKNOWLEDGEMENTS

Research partially supported by M.U.R.S.T., Rome, Italy, and by L. Bocconi University, Milan, Italy.

REFERENCES

CLAYTON, D.G. (1991), A Monte Carlo method for Bayesian inference in frailty models. *Biometrics* **47**, 467–485.

CONSONNI, G. and VERONESE, P. (1995), A Bayesian method for combining results from several binomial experiments. *Journal of the American Statistical Association,* 90, 935–944.

KALBFLEISCH, J.D. and PRENTICE, R.L. (1980), *The statistical analysis of failure time data.* Wiley, New York.

RAFTERY, A.E., MADIGAN, D. and VOLINSKI, C.T. (1995), Accounting for model uncertainty in survival analysis improves predictive performance. To appear in: *Bayesian Statistics 5,* Oxford University Press, Oxford.

PRENTICE, R.L. (1973), Exponential survivals with censoring and explanatory variables. *Biometrika* **60**, 279–288.

UN MODELLO GERARCHICO PER IDENTIFICARE FATTORI PROGNOSTICI NELL'ANALISI DELLA SOPRAVVIVENZA

RIASSUNTO

Il modello di Cox rappresenta una delle metodologie statistiche più diffuse per l'analisi dei dati di sopravvivenza. Tuttavia, tale modello può risultare inadeguato nella fase preliminare, quando molteplici fattori esplicativi debbono essere specificati e selezionati. Nel presente lavoro suggeriamo una metodologia alternativa, meno strutturata, per l'identificazione dei fattori prognostici.

Parole chiave: *modelli gerarchici partizionati, probabilità a posteriori, analisi della sopravvivenza.*