on semi-parametric survival analysis using SAS

a Bayesian Piecewise Exponential Model for assessing risk in subjects affected by sarcoma

Giuseppe Marano, Patrizia Boracchi and Elia Biganzoli

Unit of Medical Statistics, Biometry and Bioinformatics, Fondazione IRCSS Istituto Nazionale Tumori di Milano,
Department of Clinical Science and Community Health, Università degli Studi di Milano

Survival Analysis

concerns the statistical assessment of the time of occurrence of specific events, such as, for example, the progression of a disease. Special methods of analysis are required because of:

❑ **censoring mechanism:** the event of interest is not observed in all sample units. Censored (incomplete) times provide partial information about the 'true' times of occurrence.

The most common is **right censoring**: the event does not occur until a certain time after which no information is available about subject status (e.g. drop out from the study, end of follow-up).

❑ **non gaussian distribution:** defined on positive (real or integer) numbers, and typically skewed to the right.

However GLM techniques can be used to fit specific Survival Models.

BASIC TARGET FUNCTIONS

Used in the analysis of a continuous time-to event (T) in general situations:

1. HAZARD FUNCTION:

$$h(t)=\lim_{\Delta t\to 0}\frac{P\{T\in[t,t+\Delta t)\mid T\geq t\}}{\Delta t}$$ instantaneous hazard

h(t) = 'probability' that, if not occurred until t, the event will occur in the immediately following instant

2. SURVIVAL FUNCTION:

$$S(t)=P\{T>t\}=1\text{-}F(t)=\exp\left(-\int_0^t h(u)du\right)$$

S(t) = probability that the event does not occur before t

Background and Aim

In bio-statistical applications non-parametric and semi-parametric methods have been preferred over parametric ones for assessing the prognostic role of clinical/biological variables: in fact parametric distributional models for times to event are usually restrictive. In particular **Proportional Hazard Models** are widely adopted due to their simplicity and flexibility. In PH models the effect of prognostic variables is represented as a multiplier of a baseline hazard function $h_0(t)$: so that the *relative hazard* of any two subject profiles does not vary with time.

$$h(t\mid \underline{x})=h_0(t)\exp(\underline{x}'\underline{\beta})$$

The most frequently used model is the **Cox Model**, in which no assumption of the functional form of $h_0(t)$ is made. However, such characteristic becomes a drawback if the interest lies on the hazard function or in predictive modeling. In the **Piecewise Exponential Model** (PE) the baseline hazard $h_0(t)$ is piecewise constant on a partition of the time axis: this specification preserves flexibility without requiring restrictive distributional assumptions. Time intervals are included as predictors in the regression model through dummy variables. Furthermore, the PE model can be estimated by GLM techniques, so that it can be easily implemented with standard statistical software. The above mentioned properties have made the PE model an appealing approach for the analysis of continuous time-to-event data.

The comparison of results among different case series are generally provided in terms of regression coefficients and/or survival function S(t) with respective interval estimates. In the classical GLM framework, interval estimates of regression coefficients are directly provided, but those of S(t) are not easy to obtain.

When prior information derived from different studies is available, Bayesian methodology allow to evaluate how the information provided by the new study modifies such prior belief. Interval estimates of S(t) are derived from its posterior density. The posterior density is a by-product of the MCMC estimation method, since S(t), being a function of Markov Chains (corresponding to regression coefficients), can be treated itself as a Markov Chain. When different priors have to be evaluated a model with non informative priors could be considered as reference.

The aim is to show how the Piecewise Exponential Model can be implemented using SAS both in frequentist and Bayesian framework. Frequentist estimates were obtaines by proc GENMOD after having replicated the dataset. Bayesian estimates were obtained with PROC PHREG. The SAS code is shown in the boxes below.

APPLICATION: BAYESIAN PIECEWISE EXPONENTIAL MODEL

SAMPLE : the piecewise model was fitted to a case series of 192 subjects in care at Istituto Nazionale dei Tumori di Milano who underwent surgical resection of primary localized disease (cfr: Ardoino *et al* (2010)). The outcome of interest is in the current analysis is overall survival. Time was measured in months.

MODEL: the model was pre-specified on the base of clinical knowledge: 1) the time axis was partitioned in 15 intervals according to the schedule of patients control visits; 2) predictors were 5 known prognostic factors: age, tumor size (continuous) histologic subtype, grading, surgical margins (categorical). A non linear effect of tumor size was modelled through a flexible polynomial (Restricted Cubic Spline with 3 knots).

METHODS: concerning the application of bayesian methods, the basis of the restricted spline were orthogonalized (Gram-Schmidt method) to avoid convergence problems. The sampling algorithm was Adaptive Rejection Metropolis Sampling (default in PROC PHREG) which is particularly efficient for log-concave likelihood, and thus for the Piecewise Exponential Model. Non informative (diffuse normal) priors were chosen for the model parameters (25). Results shown below were obtained from a chain of length 6000, with burn-in 1000 and thinning by 2.

CODE FOR GLM ESTIMATION

```
PROC GENMOD data=sarcomas_long;
MODEL event = dummy1-dummy15 age size size2
              hist2 hist3 hist4 hist9 grad2 grad3 marg1 / NOINT
      DIST = POISSON LINK = LOG OFFSET = logtime ;
RUN;
```

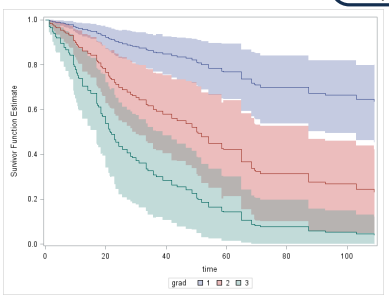
CODE FOR MCMC ESTIMATION

```
PROC PHREG data=sarcomas;
CLASS hist(REF='1') grad(REF='1') marg(REF='0');
MODEL time*event(0) = age size size2 hist grad marg;
BAYES  SEED=1973
   SAMPLING=ARMS NBI=1000 NMC=6000 THIN=2  CPRIOR=NORMAL
   PIECEWISE=LOGHAZARD( PRIOR=NORMAL
      INTERVAL= (0,4,8,12,16,20,24,30,36,42,48,54,60,72,96) )
   OUTPOST = postsample;
BASELINE OUT=base COVARIATES=covar SURVIVAL=_ALL_;
RUN;
```

Parameter	GLM		MCMC	
	Estimate	Standard Deviation	Posterior Mean	Standard Deviation
Alpha1	-5.6831	0.5377	-5.7878	0.5412
Alpha2	-5.8976	0.5856	-6.0280	0.5978
Alpha3	-4.9754	0.4839	-5.0701	0.4824
Alpha4	-5.9597	0.6824	-6.1792	0.7106
Alpha5	-4.8227	0.5074	-4.9160	0.5022
Alpha6	-4.8585	0.5226	-4.9616	0.5255
Alpha7	-5.4660	0.5720	-5.5901	0.5874
Alpha8	-5.2488	0.5682	-5.3774	0.5930
Alpha9	-5.4810	0.6806	-5.7076	0.7170
Alpha10	-5.7564	0.7970	-6.0237	0.8492
Alpha11	-4.5893	0.5748	-4.7271	0.5884
Alpha12	-5.3210	0.7957	-5.6381	0.8725
Alpha13	-4.6401	0.5415	-4.7470	0.5546
Alpha14	-5.9504	0.7805	-6.2384	0.8763
Alpha15	-5.6122	0.7850	-5.8878	0.8559

Parameter	GLM		MCMC	
	Estimate	Standard Deviation	Posterior Mean	Standard Deviation
age	0.0186	0.0098	0.0188	0.0099
size	0.0116	0.0112	0.0109	0.0113
size2	-0.1258	0.0536	-0.1307	0.0536
hist2	-0.4768	0.3773	-0.5125	0.3813
hist3	-1.7972	1.0295	-2.3850	1.2663
hist4	-0.2246	0.4672	-0.2794	0.4774
hist9	0.4631	0.4077	0.4157	0.4180
grad2	1.2129	0.3913	1.2233	0.3775
grad3	2.0570	0.3776	2.0864	0.3670
marg1	1.4250	0.3659	1.3896	0.3744

Estimated of model parameters: basal hazard (left) and regression coefficients (right)
Frequentist and Bayesian estimates and standard deviations show a good overall agreement.



Bayesian estimates of the survival function with Highest Posterior Density intervals.
The curves shown in the figure were calculated for specific predictors values: age= 60 y, tumor size = 20 cm, margin resection = microscopic, histological type = Liposarcoma, grading = 1,2,3.

CONCLUSIONS

Proc GENMOD provided estimates of regression coefficents of the piecewise exponential model (frequentist approach), but required to modify the dataset. S(t) and confidence intervals are not provided.

The Bayesian model was fitted without incurring in convergence problems: standard diagnostic techniques showed appreciable mixing properties of the Markov Chains, with autocorrelations rapidly approaching zero. This was due to the efficiency the ARMS algorithm implemented in PROC PHREG. The analysis required few efforts in programming and few computational resources.

In the current application results of the two estimation approaches are showed in terms of regression coefficients and their standard errors. The two approaches described above provided very similar results.

The facilities added to PROC PHREG in recent versions of the SAS System (9.2 and later) seem to provide powerful and user-friendly tools, adequate for performing a standard Bayesian analysis in the context of semi-parametric survival modeling.

This work was funded by the Italian Association for Cancer Research (AIRC)
IG 2012 rif: 13420, "Statistical Tools for Prognosis and Prediction in Cancer: Assessments and Application to a Sarcoma Case Series". Giuseppe Marano was a fellow of AIRC.

REFERENCES

✓ Ardoino, I., Miceli, R., Berselli, M., Mariani, L., Biganzoli, E. M., Fiore, M., Collini, P., Stacchiotti, S., Casali, P. G., Gronchi, A. : Histology - specific nomogram for primary retroperitoneal soft tissue sarcoma. Cancer, 2010. 116(10), 2429-2436.
✓ Gardiner, J. C. : Survival analysis: overview of parametric, nonparametric and semiparametric approaches and new developments. In: SAS Global Forum 2010. Statistics and Data Analysis. 2010.