Multivariable Cox Proportional Hazard Model by SAS PHREG & Validation by Bootstrapping Using SAS Randomizers With Applications Involving Electrocardiology and Organ Transplants

Sudhanshu K. Ghoshal, Ischemia Research Institute, San Francisco, CA

ABSTRACT:

The paper presents statistical procedures of multivariable model building for survival analysis of patients undergoing surgery with many preoperative risk factors. In the first example we have added electrocardiologic risk factors to the traditional clinical and demographic risk factors. All analyses were performed on SAS 6.12. The procedure described here mainly concentrates on Cox's regression analysis with risk factors assumed to be constant over time. In the last section a more generalized version of Cox's proportional hazard model is presented. Some possible applications have been presented.

1.0 INTRODUCTION:

The paper mainly discusses procedures for the analysis of survival data for patients undergoing non-cardiac surgery (example-1) and heart transplants (example-2). Our analysis included Cox's Multivariate Proportional Hazard Models (SAS PHREG) with stepwise selection process. The models were validated by bootstrapping based on Efron's technique and the samples were generated by SAS Randomizer. The next two sections contain brief description of the types of the proportional hazard models used in our study.

2.0 COX'S MODEL WITH RISK FACTORS ASSUMED TO BE CONSTANT OVER TIME

Suppose the survival times of n individuals are available, where 'r' of these are deaths and remaining n- r are living. If we assume that there are p risk factors (explanatory variables) so that

$$x=(x_1 , x_2, .. x_p)',$$

Let $h_0(t)$ be the hazard function for an individual for whom the values of all the risk factor that make up the vector x are zero. The function $h_0(t)$ is called the baseline hazard function. The hazard function of the i-th individual can be written as

$$h_i(t) = \psi(x_i)h_0(t),$$

where $\psi(x_i)$ is a function of the values of the vector of explanatory variables of the i-th individual. The function $\psi()$ may be interpreted as the hazard at time t for an individual whose vector of explanatory variables is x_i, relative to the hazard for an individual for whom x=0.

Since hazard function is nonnegative, $\psi(x_i)$ may be expressed as

$$\psi(x_i) = \exp(\beta' x_i)$$

where

$$\eta_i = \beta' x_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \quad +\beta_p x_{1p} \text{ for } i= 1,2,3,...n,$$

This is the most commonly used model for the analysis of survival data.

So the general proportional hazard model becomes

$$h_i(t) = \exp(\beta' x_i)h_0(t)$$

All the risk factors are assumed to be constant over time

$\eta_i=\beta_1 x_{1i} + \beta_2 x_{2i} + \quad +\beta_p x_{1p}$ is called the linear component of the model. It is also called the risk score or prognostic index for the i-th individual

Let us consider a case where the data consists of n observed survival times, denoted by $t_1, t_2, \ldots t_n$ and δ_i is a censoring function which is zero if the i-th survival time t_i is right-censored and unity otherwise.
The β coefficients in the proportional hazard model may be estimated by the method of maximum likelihood. Using Cox's analysis(1972) the corresponding likelihood function for the proportional hazard model is given by

$$\prod_{j=1}^{r} \{\exp\beta' x_{(j)} / (\sum_{l \in R(t_{(j)})} \exp(\beta' x_l))\}$$

where $R(t_i)$ is the risk set at time t_i.

The corresponding log-likelihood function is given by

$$\sum_{i=1}^{n} \delta_i \{\beta' x_i - \log \sum_{l \in R(t_i)} \exp(\beta' x_l)\}$$

The maximum likelihood estimates of the β-parameters in the proportional hazard model can be found by maximizing this log-likelihood function using numerical methods.

A brief discussions are presented in example-1. The cases of ties are handled by Efron's technique in PROC PHREG.

3.0 COX'S GENERALIZED REGRESSION ANALYSIS WITH TIME DEPENDENT RISK FACTORS:
Cox's generalized hazard function model may be explained as follows: In this case
$x(t)=(x_1(t), x_2(t) \ldots x_p(t))$ ' the hazard function for the i-th individual is
$h_i(t)=\exp(\beta' x_i(t))h_0(t)$
where i= 1,2,3, ...n, and
$\beta' x_i(t)=\beta_1 x_{1i}(t)+\beta_2 x_{2i}(t)+\ldots+\beta_p x_{pi}(t)$
is the value of the componeht of the linear predictor of the model for the i-th individual.$h_0(t)$ is the baseline hazard function for an individual for whom all the variables are zero at the time origin and remain at this same value through time. Here $x_{ji}(t)$ is the value of the j-th risk factor at time t for the i-th individual, some or all of the risk factors may be time dependent.

Integrating the hazard function the survivor function for the i-th individual is given by

$$S_i(t) = \exp\{-\int_0^t \exp(\sum_{j-1}^{p} \beta_j x_{ji}(u))h_0(u)du\}$$

The corresponding partial log-likelihood function may be written as

$$\sum_{i=1}^{n} \delta_j \{\sum_{j=1}^{p} \beta_j x_{ji}(t_i) - \log \sum_{i \in R(t_i)} \exp(\sum_{j=1}^{p} \beta_j x_{ji}(t_i))\}$$

Particular cases of this model has been used in organ transplant studies.

A brief discussions are presented in example-2.

4.O VALIDATION BY BOOTSTRAPPING:

It is important to validate the models whenever possible

Normally when we have large number of observations, the data may be split into two parts:
-training set -validation set

to perform cross validation.

However when we have limited amount of data we cannot split the data into two parts. However,it is possible to validate the model developed by bootstrapping. This is based on the technique developed by Efron. The concept may be described as follows: Suppose we have a random sample $x = (x_1, x_2, .. x_p)$ from an unknown probability distribution F .We wish to estimate a parameter of interest

$\Theta = t(F)$

on the basis of x. For this we estimate $\hat{\Theta} = s(x)$ from x.

Problem is to determine how accurate is $\hat{\Theta}$. The bias of the estimate is given by

$bias(\hat{\Theta}, \Theta) = E_F[s(x)] - t(F)$

Bootstrap methods depends on the notion of a bootstrap sample. A bootstrap sample

$x_1^*, x_2^* \ldots\ldots x_p^*$

is defined to be a random sample of size n drawn with replacement from a population of n subjects $(x_1, x_2 .. x_p)$ So the bootstrap set $(x_1^*, x_2^* \cdots x_p^*)$ consists of a member of the original data set some appearing zero times, some appearing once, some appearing twice etc. Corresponding to a bootstrap data set x^* there is a bootstrap replication of $\hat{\Theta}^* = s(x^*)$

The quantity $s(x^*)$ is the result of applying the same function s(.) to as x^* as was applied to x.

For example if s(x) is the sample mean $\bar{x}$ then $s(x^*)$ is the mean of the corresponding bootstrap data set

5.0 THE ALGORITHM FOR ESTIMATING STANDARD ERROR

1. Select B independent bootstrap samples $x^{*1}, x^{*2} \ldots\ldots x^{*B}$ each consisting of n data values drawn with replacement from $x = (x_1, x_2 .. x_p)$

According to Efron B will normally be in the range of 25-200.

2. Evaluate the bootstrap replication corresponding to each bootstrap sample:

$\hat{\Theta}(b) = s(x^{*b})$ where b=1,2, B

3.Estimate the standard error $S_e(\Theta)$ by the sample standard deviation of the B replications where

$$\hat{S}_{eB} = \{\sum_{b=1}^{B} [\hat{\Theta}^*(b) - \hat{\Theta}^*(.)]^2 / (B-1)\}^{1/2}$$

where $\hat{\Theta}^*(.) = \sum_{b=1}^{B} \hat{\Theta}^*(b) / B$

The estimate of the bias is

$$\hat{bias}_B = \hat{\Theta}(.) - t(\hat{F})$$

6.0 BOOTSTRAP ESTIMATE OF PREDICTION ERROR:

Select B independent bootstrap samples $x^{*1}, x^{*2}, \ldots x^{*p}$ as before. Apply logistic regression or Cox's regression with stepwise selection to find a best fit model to each of these bootstrap samples.
The model is applied to the original data the prediction error err_orig.

The model is applied to the x*i bootstrap sample data the apparent error is err_boot.
optimism = err_orig - err_boot

The overall estimate of optimism is the average of the B differences. Once an estimate of optimism is obtained, it is added to the apparent error rate to obtain an improved estimate of the prediction error.
Example: Suppose we wish to estimate the expected value of the Sumers' D rank correlation coefficient between predicted and observed survival time. This may be done as follows:

a. Develop the model using original population with say logistic regression stepwise selection procedure. Let Dapp denote the apparent D from this model.
2. develop the stepwise model on one of the B samples. The apparent D computed on this bootstrap model is Dboot.
3. Apply this model to the original data and calculate D on the original data, call it Dorig
4. The optimism in the fit is
diff=Dorig - Dboot

5. The step is repeated B times.

6. The average of diff is the average optimism call it opt_av
The bootstrap corrected performance of the original stepwise model = Dapp + opt_av
This is a nearly unbiased estimate of the expected value of the external predictive discrimination of the process which generated Dapp
Similar strategy is applied for estimating the prediction error at time t in a survival model. Instead of computing Sumers D we compute the statistic D_Stat = difference between mean predicted 2-year survival probability and Kaplan-Meier 2-year survival estimate

7.0 Example 1:

In this example we have described the statistical procedure of multivariable model building in a survival analysis of 336 surgery patients having many preoperative risk factors. The risk factors include demographic, traditional clinical and electrocardiologic variables. The procedure Proc Corr was employed to determine the correlation between the risk factors and interaction terms were added whenever necessary. The procedure described here concentrates on Cox's
regression analysis with risk factors not varying with time. In the process of determining associations of various risk factors to patient mortality univariate LOGISTIC REGRESSION was applied. Our next step was to determine the best main effect multivariable model by fitting a subset of predictors (with p value <0.1).We have employed Proc PHREG with stepwise selection process. Models were developed for the different periods of follow up (inhospital, 1 year, 2 year, etc). In the final models predictors with multivariable two-tailed p <.05 were retained. For each period of survivability prognostic risk scores were computed from the Cox regression coefficients. The distribution of prognostic scores was divided into three risk categories - low, moderate and high. The upper threshold was set equal to the average of medium and maximum values of the prognostic scores while the lower threshold was set equal to the average of mean and lower quartile of the prognostic scores. Kaplan-Meier survival curves, probability of survival were computed for the three risk categories and each year (period) of follow up. The models were validated by bootstrapping based on the generalization

of Efron's technique using SAS randomizer.Cases of ties were handled by Efron's technique.

BOOTSTRAPPING

200 samples were created from the original populations of the patients employing SAS randomizer RANUNI(seed, x). The estimated standard error of the mean of survival estimates of the samples was small for all the risk groups .

Example -2:

Crowley and Hu have shown how time dependent risk factors can be used in organ transplantation. Here the risk factor $x_1(t)$ takes the value 0 if the patient has not received a transplant at time t and unity otherwise, the hazard of death for the i-th individual at time t is given by

$h_i(t) = \exp\{\eta_i + \beta_1 x_{1i}(t)\}h_0(t)$

η_i is a linear combination of other non-time-dependent explanatory variables whose values have been recorded at the time origin for the i-th individual, x_{1i} is the value of x_1 for the individual at time t. Cox and Oakes (1984) suggested an improvement to the hazard model. Details are omitted for lack of space.

Acknowledgments:

The views and results presented here are the personal opinion of the author and Ischemia Research Institute is not anyway responsible for them.

Dr.Sudhanshu K. Ghoshal Ph.D, D.Sc
Senior Research Scientist
Ischemia Research Institute
San Frncisco, CA 94134-3306

9.0 References:

Collett, D (1994) : Modeling Survival Data in Medical Research, London: Chapman & Hall

Cox, D.R (1972) : Regression Models and Life Tables (with Discussion), Journal of the Royal Statistical Society, B34, 187-220.

Cox, D.R and Oakes, D (1984) : Analysis of Survival Data, Chapman and Hall, London
Crowley, J and Hu, M.(1977) Covariance Analysis of heart transplant survival data. Journal of the American Statistical Association, 78, 277-81

Efron B (1979) : Bootstrap Methods : Another Look at the Jackknife: Annals of Statistics 7: 1-26

Efron B (1981a) : Censored data and the bootstrap J of American Statistical Asso, 76:312-319

Efron B (1981b): Nonparametric standard error and confidence intervals: Canadian Journal of
Stat,9, :139-172

Efron B and Tibshirani R(1986): Bootstrap Methods for standard errors : SIAM, Philadelpia

Example-1: Results:
Model:A. Cox's Proportional Hazard Model for 1 year survivability

No of Risk Factor	Hazard Ratio	p-value	Coefficient Parameter
1. SDNN < 50 msec	2.5	.0017	.919541
2. ASA Physical Status >=4	2.1	.0181	.727915
3.Ventricular Tachycardia	3.3	.0039	1.186987
4.Index Surgery for cancer (VT)	5.1	.0001	1.622711
5.Moderate or Severe Limitation of Activity	2.9	.0002	1.058695
6.Estimated Creatinine Clearance < 0.83	2.7	.0005	.991559

SDNN = standard deviation of normal to normal QRS intervals,
ASA = American Society of Anesthesiologists

In previous studies Clinicians found demographic & clinical variables were associative with mortality. It is nice to find the Electrocardiologic variables No 1 & 3 are also associative and included in the model. A. We used SAS randomizer to generate 200 random samples out of the original population of patients (336).. On univariate analysis of on the 200 samples we found the following

	No of times significant out of 200	%	
1. SDNN < 50 msec	159/200	80%	
2. ASA Physical Status >=4	106/200		53%
3. Ventricular Tachycardia	154/200		77%
4. Index Surgery for cancer	163/200		82%
5. Moderate or Severe Limitation of Activity	163/200	82%	
6. Estimated Creatinine Clearance < 0.83	180/200	90%	
7. Couplet	82/200	41%	
8. Couplet or VT	67/200	34%	
9. Using Bronchodilator Medication	111/200	56%	
10. Histchf	136/200	68%	

etc etc. Rest of them were lot less associative with mortality.
Broncho & histchf were not significant in the original population & was not selected (forward selection) in the final model. Their p-values were > .1. We retained risk factors with p-values < .1 in univariate analysis & < .o5 in the final model. Fig -1 shows survivability with hi, moderate and low prognostic scores. Fig -2 One to five year survivability with respect to risk scores in general. Fig 3 shows survivability with or without electrocardiologic variables (SDNN & VT).