Paper 236-2009

Fitting Cox Model Using PROC PHREG and Beyond in SAS

Lea Liu, Sandy Forman, Bruce Barton
Maryland Medical Research Institute, Baltimore, Maryland, USA

Abstract

Cox proportional hazard model has been widely used for survival analysis in many areas in investigating time-to-event data. PROC PHREG in SAS has been a powerful tool used for construction of a Cox model. In addition to the STATEMENTs and OPTIONs within PHREG that have already provided the most demanded output, a little more effort on data manipulation would accomplish the calculations and generate output for some new statistical approaches. This paper presents: 1) Using overall C-index as a measure of discrimination for model validation; 2) Calculating adjusted survival rate using corrected group prognosis method; 3) Presenting the effect of a continuous covariate on estimated survival. The core SAS program code and MACRO used for implementation are included.

Introduction

Cox proportional hazard model has been widely used for survival analysis in many discipline areas in investigating time-to-event data. New statistical methods and approaches on this subject have been updated consistently. PROC PHREG in SAS has been a powerful tool used for construction of a Cox model. The STATEMENTs and OPTIONs within PROC PHREG have provided the most demanded output. However, there is a lag time for SAS to update the code to respond to the new methods. To meet the needs in daily work, we have created some SAS programs and a MACRO for the following three new statistical approaches: 1) Using overall C-index as a measure of discrimination for model validation; 2) Calculating adjusted survival rate using corrected group prognosis method; 3) Presenting the effect of a continuous covariate on estimated survival. These methods have been used in many medical research articles in recent years.

A model construction usually starts with fitting a ‘full’ model including all candidate predictors. Then a preferred approach for selecting variables will be used to obtain a ‘final model’ which is more stable and retains the predictive power with reduced number of variables that make significant contribution to the outcome. The three new statistical approaches we introduced in this paper all apply on the final model.

The examples in this paper use the same data set from a clinical study on heart diseases.

Model validation using overall C as a measure of discrimination

Discrimination is one of model validation process that tests the ability of a predictive model to separate those who develop event from those who do not. One of the most popular measures of discrimination is ROC curve. SAS has options for generating classification table and ROC curve in PROC LOGISTIC. However measurement of predictive accuracy can be more complex for survival analysis in the presence of censoring. C-index introduced by Harrell (Ref. 1, 1996) as a natural extension of the ROC curve is an easily interpretable measure of predictive discrimination. Later, Pencina and D’Agostino (Ref. 2, 2004) developed the overall C index as a parameter describing the performance of a given model applied to the population under consideration and discuss the statistic used as its sample estimate.

C-index for the survival analysis model is defined as the probability of concordance given that the pairs considered are usable in which at least one had an event. It can be interpreted as the probability that a subject from the event group has a higher predicted probability of having an event than a subject from the non-event group.

In constructing C-index, we can use only usable pairs. This results in either event vs. event or event vs. non-event comparison.

Example 1: we use the factors (MI history (mihx), Diabetes (diabhx), Low Ejection Fraction (lowef)) to explain the composite end point of death, re-infarction, or class IV heart failure (combfv).

First, re-run the final model using PROC PHREG with OUTPUT statement to create dataset that contains subject-id, observed survival time and survival function estimate for each individual. Then create a dataset 'evtset' including only the subject who had event.

```
proc phreg data=sample;
  id idn;
  model combdays*combfv(0)=mihx diabhx lowef;
  output out=obs survival=surv;
run;
data evtset; set obs; if combfv=1; rename idn=idn_j surv=y_j combdays=x_j;
keep idn surv combdays; run;
```

Secondly, construct all usable pairs and create variable for concordance. PROC SQL creates dataset that including all usable pairs by a Cartesian join. Dataset 'concord' includes a new variable concord that identifies if the pair is concordant or not.

```
proc sql;
      create table allset as
      select idn_j, y_j, x_j, idn as idn_i, surv as y_i, combdays as x_i
      from evtset, obs
      where idn_j<>idn;
quit;
data concord;
      set allset;
      if (x_i<x_j and y_i>y_j) or (x_i>x_j and y_i<y_j) then concord=1;
      else concord=0;
run;
```

Denote the actual survival times of subjects by X1, X2, …, the predicted probabilities of survival by Y1, Y2, …, a concordant pair is when Xi<Xj and Yi<Yj or Xi>Xj and Yi>Yj. If the inequalities go in the opposite direction, i.e.: Xi<Xj and Yi>Yj or Xi>Xj and Yi<Yj, then a pair is said to be discordant.

The following code performs the calculation of the C-index and 95%CI using the estimated probabilities of concordance and discordance proposed by Pencina and D'Agostino.

```
data _null_;
     set concord end=eof;
     retain nch ndh;
     if _N_=1 then do;
       nch=0;
       ndh=0;
       end;
     if concord=1 then nch+1;
     if concord=0 then ndh+1;
     if eof=1 then do;
       call symput('ch',trim(left(nch)));
       call symput('dh',trim(left(ndh)));
       call symput('uspairs',trim(left(_n_)));
       end;
```

```
run;

data _null_;
     set sample end=eof;
     if eof=1 then call symput('totobs',trim(left(_n_)));
run;

%put &ch &dh &uspairs &totobs;

data calculat;
     ch=input("&ch",12.0);
     dh=input("&dh",12.0);
     uspairs=input("&uspairs",12.0);
     totobs=input("&totobs",10.0);
     pc=ch/(totobs*(totobs-1));
     pd=dh/(totobs*(totobs-1));
     c_hat=pc/(pc+pd);

     w=(2*1.96**2)/(totobs*(pc+pd));
     low_ci_w=((w+2*c_hat)/(2*(1+w)))-(sqrt((w**2+4*w*c_hat*(1-
     c_hat))/(2*(1+w))));
     upper_ci_w=((w+2*c_hat)/(2*(1+w)))+(sqrt((w**2+4*w*c_hat*(1-
     c_hat))/(2*(1+w))));
run;
```

In this example, there are total 2185 subjects with 310 events. The final model:

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
mihx	1	0.40059	0.15242	6.9076	0.0086	1.493
diabhx	1	0.63880	0.12146	27.6618	<.0001	1.894
lowef	1	0.72633	0.12439	34.0949	<.0001	2.067

Table 1. The C-index and 95% CI of the model:

Number of usable pairs	*Number of concordant pairs*	*Number of discordant pairs*	*C-Index*	*Lower 95%CL*	*Upper 95%CL*
677040	519081	157959	0.7667	0.6656	0.8549

Calculating adjusted survival rates using corrected group prognosis method.

Calculation of adjusted survival curve or estimates of risk-adjusted survival from proportional hazard model is a common task for survival analysis. In SAS, PROC PHREG with BASELINE COVARIATES statement generates the estimation of adjusted survival rates using mean of covariates method. This method applies average value of covariate on the model and gets the average hazard (the hazard for the average individual) which is not the same as the average survival estimated form a heterogeneous group of individuals. Because of concerns for this method, several other approaches have been proposed to overcome the shortcomings. One of the better alternatives is corrected group prognosis method (Ref. 3 and Ref. 4). It calculates the survival curve for each individual using fitted Cox model, the average survival is then calculated as a weighted average of the individual survival curve. Programs for its application were also made available on website or articles. However, they are not friendly enough for user to pick-up. A SAS MACRO presented with this paper would make the process much easier.

Example 2: Calculate adjusted 5-years survival based on a developed Cox model for diabetic and non-diabetic patients with control for age (age65), ejection fraction (lowef) and congestive heart failure (chfhx). The program (CGPM_adj.sas) used for this example is included in Appendix. You can simply replace the parameters at the beginning of the program, then highlight the rest of the program and run for your data.

The output of the MACRO includes: (1) Summary table for the model shown as Table 2 below. (2) Summary table shown as Table 3 for the calculated unadjusted and adjusted survival estimates and death (1-survival) at the time of last event observed in 5-years observation. (3) Graph with unadjusted and adjusted survival curve by group shown as Figure1.

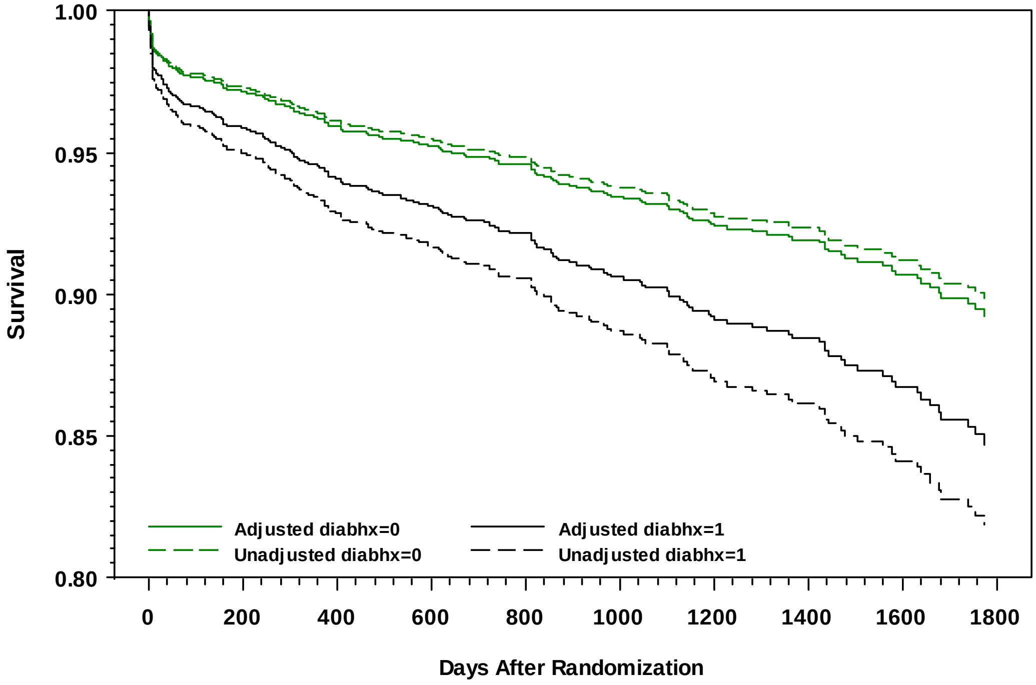
Table 2. Analysis of Maximum Likelihood Estimates (alpha=0.05 for CI)

Variable	*DF*	*Estimate*	*StdErr*	*ChiSq*	*ProbChiSq*	*HazardRatio*	*HRLowerCL*	*HRUpperCL*
diabhx	1	0.38650	0.16560	5.4475	0.0196	1.472	1.064	2.036
age65	1	0.49613	0.15036	10.8880	0.0010	1.642	1.223	2.205
lowef	1	0.76629	0.16773	20.8711	<.0001	2.152	1.549	2.989
chfhx	1	1.28559	0.24918	26.6177	<.0001	3.617	2.219	5.894

Table 3. Estimated survival and event rates at the time of last event observed

group	*survival*	*1-survival*
Unadjusted diabhx=0	0.89852	0.10148
Adjusted diabhx=0	0.89265	0.10735
Unadjusted diabhx=1	0.81846	0.18154
Adjusted diabhx=1	0.84764	0.15236

Figure 1. Adjusted Survival Curves (solid line) by Corrected Group Prognosis Method Compared with Unadjusted Survival Curves (dash line).



Presenting the effect of a continuous covariate on estimated survival.

On SUGI28, Allmer (Ref. 5, 2000) proposed an approach to visualizing the effect of a continuous covariate on estimated survival. The method uses PROC PHREG with BASELINE statement to output estimated survival function for each combination of the explanatory variable values present in the dataset; then, select the estimated survival probability at a desired time point for each subject in the study; then plot the estimated survival probability against the selected continuous variable. A smoothing method is then used to provide an estimate of the impact of the covariate on the survival rate at the time point of interest over the range of covariate values.

We here propose an enhanced plot using the same technique by adding estimated confidence interval generated by BASELINE statement. This type of plot is more informative and more useful for investigating the effect of a continuous variable on response variables. This approach can also be used with adjustment for other considered covariates.

Example 3: Present the effect of age as a continuous variable on the composite primary outcome control for gender using the Cox model (shown on Table 4).

Table 4. Cox model for examine the effect of age on composite primary outcome control for gender in study.

Analysis of Maximum Likelihood Estimates

Variable	*DF*	*Parameter Estimate*	*Standard Error*	*Chi-Square*	*Pr > ChiSq*	*Hazard Ratio*
age	1	0.01862	0.00540	11.8847	0.0006	1.019
male	1	-0.28396	0.12996	4.7743	0.0289	0.753

The lines in Figure2 were obtained by using SAS PHREG with BASELINE statement to calculate an estimated survival probability and 95% confidence limits for each combination of gender and age (years) values presented in the dataset, then selected only the 5-years survival probability for each person in the study, then calculate the estimated 5-years primary outcome rate as 1- estimated survival probability for each individual, then plotting the estimated 5-years primary outcome rates with 95%CI against the age values by gender group using PROC GPLOT. The time point for 5-years was represented by the time of last event observed that was 1683 days for males and 1478 for females in the database. The core of SAS code is shown as below:

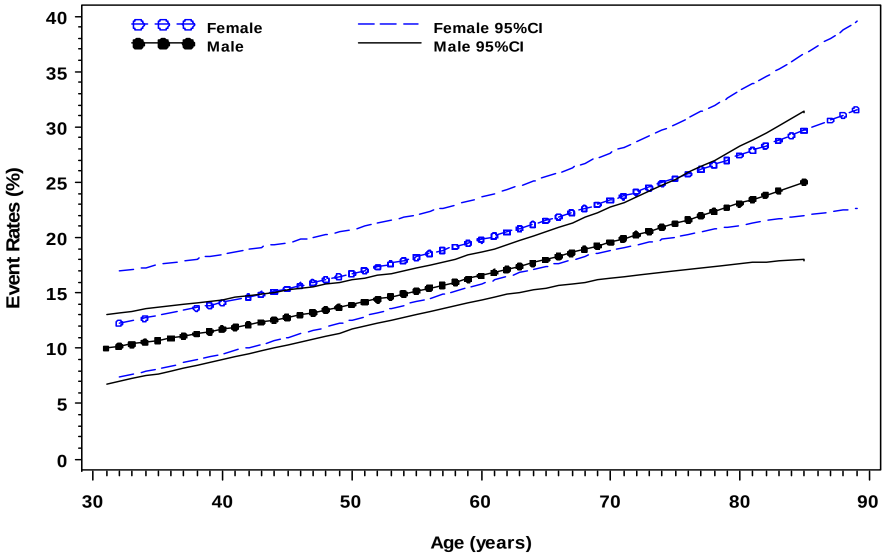
```
proc phreg data=sample;
model combdays*combfv(0)=age male;
baseline out=yrout lower=low_ci upper=up_ci survival=surv
covariates=sample/alpha=0.05 nomean;
run;

proc sort data=sample out=maxday (where=(combfv=1)
               keep=idn male combfv combdays);
      by male combdays;
run;

data _null_;
      set maxday;
      by male;
      if male=0 and last.male then call symput('lastday_f',combdays);
      if male=1 and last.male then call symput('lastday_m',combdays);
run;
%put &lastday_f &lastday_m;
```

```
data yrest;
      set yrout;
      evtrate=(1-surv)*100;
      evtlow=(1-low_ci)*100;
      evtup=(1-up_ci)*100;
      if combdays=&lastday_m and male=1 then output;
      if combdays=&lastday_f and male=0 then output;
run;
```

Figure 2. Visualizing the impact of age (years) as continuous covariate on estimated 5-years primary outcome with 95% confidence interval by gender group.



Conclusion

The three approaches introduced in this paper are commonly demanded in survival analysis for bio-medical research. The programs proposed here are useful tools for the implementation of the approaches in addition to PROC PHREG. Intermediate level of SAS programming and knowledge on survival analysis using Cox model are needed to understand the process. The programs were developed and implemented in SAS 9.1.3 Service Pack 4, XP-PRO Platform.

References

1. Harrell FE et al. Tutorial in Biostatistics: Multivariable prognosis model: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine 1996, 15:361-387.
2. Pencina MJ and D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Statistics in Medicine 2004, 23:2109-2123.
3. Nieto FJ and Coresh J. Adjusting survival curves for confounders: A review and a new method. American Journal of Epidemiology 1996, 143(10): 1059-1068.
4. Ghali WA et al. Comparison of 2 methods for calculating adjusted survival curves form proportional hazard models. JAMA 2001, 286(12): 1494-1497.
5. Allmer C et al. An approach to displaying predicted survival data based on the level of a continuous covariate. SAS online proceedings 2000, SUGI28, paper 201-28.

Appendix (SAS code)

```
*****************************************************************************;
* Program:  CGPM_adj.sas;
* Purpose:  Calculate the unadjusted and adjusted survival rates and plot the
*           survival curves using corrected group prognosis method;
*****************************************************************************;

*----------------------------------------------------------------------------;
* 1. All variables used in the model have to be binary variable.
*     Categorical and numeric variables need to be re-coded as dummy
*      variable to meet this requirement;
* 2. The maximum number of variables in the model is limited to 20;
*     (not including the control variable). But this number can be
*     easily expanded as needed.;
* 3. The analysis dataset for modeling was assign to be in SAS
*     default library WORK. ;
*----------------------------------------------------------------------------;
*----------------------------------------------------------------------------;
* put your specifications on the right of the equal sign (=);
*----------------------------------------------------------------------------;
*specify the output location;
%let outputpath=L:\SGF2009\Examples;
*specify the file name for the graph, no extension name needed;
%let figname=Adjdiab_fig;
*specify the file name for the summary table, no extension name needed;
%let tbname=adjdiab_summary;
*specify the dataset name that is in your SAS WORK library for the Cox model;
%let dsn=sample;
*list all covariate names, not include the control variable;
%let covars= age65 lowef chfhx;
*assign the name of control variable;
%let ctrlvar= diabhx;
*assign the name of event variable;
%let outcom=deathfv;
*assign the value of censoring for event variable;
%let cnsrval=0;
*assign the name of time variable;
%let timevar=deathdays;
*specify the alpha level for confidence interval of HR;
%let rlalpha=0.05;
*----------------------------------------------------------------------------;
* Calculate unadjusted survival rate;
*----------------------------------------------------------------------------;
*create dataset for controlled variable;
proc sql;
      create table ctrlset as
      select distinct &ctrlvar
      from &dsn
      order by &ctrlvar;
quit;
*get the unadjusted survival rates for controlled variable;
proc phreg data=&dsn noprint;
      model &timevar * &outcom (&cnsrval) = &ctrlvar;
      baseline covariates=ctrlset out=unadjset survival=survival / nomean;
run;
```

```
*----------------------------------------------------------------------------;
* Calculate adjusted survival rates using corrected group prognosis method;
*----------------------------------------------------------------------------;
*create identifier set for all possible combinations of covariates;
data _null_;
      cnt=1+count("&covars",' ');
      call symputx('varcnt', cnt);
run;
%put &varcnt;

%macro idx;
      %let i=1;
      %do %until(%scan(&covars,&i,' ')=%str());
      %local varnam;
            %let varnam=%scan(&covars,&i,' ');
            xp=&i+1;
            if &varnam=1 then substr(idx,xp,1)=1;
      %let i=%eval(&i+1);
      %end;
%mend;
%macro idxrv;
      %let i=1;
      %do %until(%scan(&covars,&i,' ')=%str());
      %local varnam;
            %let varnam=%scan(&covars,&i,' ');
            xp=&i+1;
            if substr(idx,xp,1)='1' then &varnam=1; else &varnam=0;
      %let i=%eval(&i+1);
      %end;
%mend;

data modelset;
      set &dsn;
      idx=left(put(10**(input("&varcnt",2.0)),21.));
      %idx;
      keep idx &ctrlvar &covars &timevar &outcom;
run;
proc freq data=modelset noprint;
      tables idx / out=idx;
run;
data popset (drop=percent xp);
      if _n_=1 then do until(last);
            set idx end=last;
            &ctrlvar=0;
            %idxrv;
            output;
            end;
      set idx;
            &ctrlvar=1;
            %idxrv;
            output;
run;
ods output censoredsummary=censum parameterEstimates=parasum;
proc phreg data=modelset;
id idx;
      model &timevar * &outcom (&cnsrval) = &ctrlvar &covars / rl
alpha=&rlalpha;
```

```
      baseline covariates=popset out=adjset0 survival=survival / nomean;
run;
ods output close;
proc sort data=popset out=count (keep=idx count);
by idx;
run;
proc sort data=adjset0;
by idx;
run;
data adjset1;
      merge adjset0 count;
      by idx;
run;
proc sort data=adjset1;
      by &ctrlvar &timevar;
run;
data adjset;
      set adjset1;
      by &ctrlvar &timevar;
            if first.&timevar then frqsum=0;
            frqsum+count;
            if first.&timevar then sursum=0;
            sursum+survival*count;
      if last.&timevar;
      adjsurv=sursum/frqsum;
      keep &ctrlvar &timevar adjsurv;
run;
data allsurv;
      length group $22;
      set adjset (in=a rename=(adjsurv=survival)) unadjset (in=b);
      if a=1 and &ctrlvar=0 then group='Adjusted'||' '||"&ctrlvar"||'=0';
     else if a=1 and &ctrlvar=1 then group='Adjusted'||' '||"&ctrlvar"||'=1';
      else if b=1 and &ctrlvar=0 then group='Unadjusted'||'
'||"&ctrlvar"||'=0';
      else if b=1 and &ctrlvar=1 then group='Unadjusted'||'
'||"&ctrlvar"||'=1';
      eventrate=1-survival;
run;
data summary;
      set allsurv;
      by group;
      if last.group;
run;
*----------------------------------------------------------------------------;
* You can change the code below to modify the title and graph axis scale etc.;
*----------------------------------------------------------------------------;
goption reset=all device=emf gsfname=grfa hsize=6.0 vsize=4 ftext='Arial/bo'
htext=11pt
            display;
filename grfa "&outputpath.\&figname..emf";

symbol1 i=steplj v=none l=1  ci=green w=1;
symbol2 i=steplj v=none l=1  ci=black w=1;
symbol3 i=steplj v=none l=3  ci=green w=1;
symbol4 i=steplj v=none l=3  ci=black w=1;
axis1 offset=(0.5 cm,0.5 cm);
```

```
axis2 label=(angle=90 j=c 'Survival') order=(0.8 to 1 by 0.05) offset=(0.0
cm,0.0 cm);
legend1 across=2
   mode=share
   position=(bottom left inside)
   label=none
   value=(j=l h=10pt)
   offset=(0.5cm, 0cm);

proc gplot data=allsurv;
   plot survival * &timevar = group / legend=legend1 haxis=axis1 vaxis=axis2;
label eventrate='Survival Rate'
      &timevar='Days After Randomization';
run;
quit;
title;
footnote;
ods rtf style=journal file="&outputpath.\&tbname..rtf"  bodytitle
startpage=no;
proc print data=parasum noobs;
title "Analysis of Maximum Likelihood Estimates (alpha=&rlalpha for CI)";
run;
proc print data=summary noobs;
var group survival eventrate;
title 'Estimated survival and event rates at the time of last event observed';
run;
ods rtf close;
```

Contact information

Lea Liu, Maryland Medical Research Institute
600 Wyndhurst Ave. Baltimore, MD 21210
(410) 435-4200, Email: lliu@mmri.org, Web: www.mmri.org

Acknowledgement (Copyright disclaimer)