MWSUG 2019 - Paper HW020

Survival Tips for Survival Analysis

Xiaoting Wu, Department of Cardiac Surgery, Michigan Medicine, Ann Arbor, MI

ABSTRACT

Survival analysis is a common type of analysis in health care field. This hand-on tour will convey some survival tips during survival analysis. We will provide an overview on the survival analysis using SAS® LIFETEST and PHREG procedures, including data preparation and visualization, variable selection, model specification, model validation and output interpretation. We will also showcase some advanced application on how to obtain prediction estimates, customize and output plots from SAS PHREG procedure.

INTRODUCTION

Survival analysis is a common statistical technique that look at the probability of an event occurrence over time [1]. For example, physicians may be interested in knowing the survival probability in long term after a medical procedure. Engineers may be interested in knowing the failure time of a machine [2].

Two pieces of information is needed in a survival analysis: 1) If the subjects have the event of interest (for example, an indicator status can be coded as 1=event; 0= no event (censoring); 2) The follow up time for each subject (specifically this is the time to events for those with events, and time to censoring for those without the events).

The concept of Censoring is crucial in survival analysis. Censoring indicates that the event time is unobserved for censored subjects. For a right censoring, the unobserved event time is only known to occur after the censoring time. Subjects can be censored due to the end of study, lost to follow up. A key assumption is that the censoring is independent and non-informative. In another word, if the censoring subjects were observed, they would have the same distribution of event time as those without censoring.

Kaplan Meier and cumulative incidence function are two methods to describe time to event data. COX proportional hazard model is popularly used to study the effect of the predictors that impact the event time. These methods will be illustrated in more details with a data example shown below.

DATA EXAMPLE

The data set is described in Fleming and Harrington (1991). The study originates from Mayo Clinic trial (1974-1984). PBC is a rare but fatal chronic liver disease of unknown origin, which leads to the destruction of interlobular bile ducts.

Here, we're interested in the death risk of PBC if patients did not receive liver transplant. Therefore, liver transplant is considered as a censoring event.

The data set PBC has 312 patients (of 418 total; 106 did not participate in trial) with treatment (DRUG). For 106 patients DRUG=. ; but, other variables were assessed.

Variables	Variable specification
id	case number
futime	number of days between registration and the earlier of death, transplantion, or study analysis time in July, 1986
status	0=alive, 1=liver transplant, 2=dead
drug	1= D-penicillamine, 2=placebo
age	age in days
sex	0=male, 1=female
ascites	presence of ascites: 0=no 1=yes
hepato	presence of hepatomegaly 0=no 1=yes
spiders	presence of spiders 0=no 1=yes
edema	presence of edema: 0=no edema and no diuretic therapy for edema; .5 = edema present without diuretics, or edema resolved by diuretics; 1 = edema despite diuretic therapy
bili	serum bilirubin in mg/dl
chol	serum cholesterol in mg/dl
albumin	albumin in gm/dl
copper	urine copper in ug/day
alk_phos	alkaline phosphatase in U/liter
sgot	SGOT in U/ml
trig	triglicerides in mg/dl
platelet	platelets per cubic ml/1000
stage	histologic stage of disease; 1, 2, 3 or 4.

Table 1. Variable specification of data example

LIFETEST PROCEDURE

SAS LIFETEST procedure can be used to perform Kaplan-Meier analysis and describe nonparametric estimate of survival functions. An example of commonly used LIFETEST code is given below.

```
proc lifetest data=pbc outsurv=km_sur2 plots=survival( cl test atrisk
(maxlen=13) nocensor ) maxtime=4000 notable ;
   time futime*status(0,1);
   strata drug;
run;
```

1.1 How to obtain survival probability for various time points- use outsurv option

Kaplan Meier estimates are step function, and calculated only at event times (ti), there being no information on events occurring at other times. That's why the survival estimators are missing in the censor time. We usually assume the same survival estimates for those censor times until we observed the next survival estimator at the next event time.

$$\widehat{Pr}[T > t_i | T \geq t_i] = \frac{Y_i - d_i}{Y_i}, for\ i = 1,2, \dots, D.$$

$$\hat{S}(t_i) = \widehat{Pr}[T > t_i | T \geq t_i]\ \widehat{Pr}[T > t_{i-1} | T \geq t_{i-1}] \dots \widehat{Pr}[T > t_2 | T \geq t_2]\ \widehat{Pr}[T > t_1 | T \geq t_1]$$

To simplify,

$$\hat{S}(t) = \begin{cases} 1\ \ if\ t < t_1 \\ \prod_{t_i \leq t} \left[1 - \frac{d_i}{Y_i}\right], if\ t_1 \leq t \end{cases}$$

The probability of an individual surviving beyond time x (experiencing the event after time x), defined as S(x)=Pr (X >x).

A fourth basic parameter of interest is the mean residual at time x. Given x is a continuous random variable then the p th quantile is found by solving $S(x^p)$=1-p. The median lifetime is the 50th percentile $x^{0.5}$, the 75 percentile lifetime is the 75th percentile $x^{0.75}$. ($S(x^{0.5})$ =1-0.5=0.5; $S(x^{0.75})$ =1-0.75=0.25).

In product limit estimate, we find the smallest time x^p for which the product limit estimator is less than or equal to 1-p. That is $\hat{X}_p = \inf\{t: \hat{S}(t) \leq 1 - p\}$

The missing 75th percentile survival time is because the maximum survival probability is obtained at time 0.34 from this study. Note that Kaplan-Meier estimator is only well defined for all time points less than the largest observation time. If the largest study time is a death, the survival curve is zero beyond this point. If the largest time point is censored, the value of S(t) beyond this point is undetermined.

Summary Statistics for Time Variable futime

Quartile Estimates

Percent	Point Estimate	95% Confidence Interval		
		Transform	[Lower	Upper)
75	.	LOGLOG	4191.00	.
50	3395.00	LOGLOG	3086.00	3839.00
25	1487.00	LOGLOG	1170.00	1925.00

Table 2. Kaplan Meier survival estimates

	futime	Censoring Flag: 0=Failed 1=Censored	Survival Distribution Function Estimate	SDF Lower 95.00% Confidence Limit	SDF Upper 95.00% Confidence Limit
265	3336	1	0.5114055615	.	.
266	3358	0	0.5000409935	0.4204226209	0.5744468843
267	3388	1	0.5000409935	.	.
268	3395	0	0.4884121332	0.4074842323	0.5643988816
269	3422	1	0.4884121332	.	.
270	3428	0	0.4764996421	0.3943142214	0.5540592284
271	3445	0	0.464587151	0.3813107797	0.5435993402

	futime	Censoring Flag: 0=Failed 1=Censored	Survival Distribution Function Estimate	SDF Lower 95.00% Confidence Limit	SDF Upper 95.00% Confidence Limit
280	3707	1	0.4387357197	.	.
281	3762	0	0.4241111958	0.3368720153	0.5085408252
282	3820	1	0.4241111958	.	.
283	3823	1	0.4241111958	.	.
284	3839	0	0.4084033737	0.3194866516	0.4951979523
285	3850	1	0.4084033737	.	.
286	3853	0	0.3920672387	0.3015971838	0.4812361361
287	3913	1	0.3920672387	.	.
288	3933	1	0.3920672387	.	.
289	3992	1	0.3920672387	.	.
290	4025	1	0.3920672387	.	.
291	4032	1	0.3920672387	.	.
292	4039	1	0.3920672387	.	.
293	4050	1	0.3920672387	.	.
294	4079	0	0.36900446	0.2741645848	0.4639077797
295	4127	1	0.36900446	.	.
296	4184	1	0.36900446	.	.
297	4190	1	0.36900446	.	.
298	4191	0	0.3406195015	0.2398267077	0.4438135304
299	4196	1	.	.	.
300	4232	1	.	.	.
301	4256	1	.	.	.
302	4365	1	.	.	.
303	4427	1	.	.	.
304	4453	1	.	.	.
305	4459	1	.	.	.
306	4467	1	.	.	.
307	4500	1	.	.	.
308	4509	1	.	.	.
309	4523	1	.	.	.
310	4556	1	.	.	.

Display 1. SAS dataset for Kaplan Meier survival estimates

1.2 How to truncate survival plots for time points with small at risk number - Use option MAXTIME to truncate graph.

To show the number of subject at risk at each time point, Option atrisk can be used in the PLOTS option. When there is few number of patients at risk, the confidence interval band is wide and the inference is usually limited. So researchers usually want to show the survival plots

with reasonable number of at risk subjects. Option MAXTIME can be used to truncate the plot. But notice this does not change the analysis.

1.3 How to obtain group effect in the Kaplan Meier method

To describe different survival for groups, statement STRATA can be used. And LIFETEST also provides three different test to test equality over the strata groups, including log-rank, Wilcoxon, and likelihood ratio test.

Log-rank test is standard method to test if the survival curve is different from two groups assuming the hazard is proportional constant over time. Proportional hazard means that the hazard ratio of the two groups maintains the same at all time points. When the two survival curves cross, it's possible that the hazard ratios change over the time. For example, DRUG 1 has higher survival probability in the earlier time, and DRUG 2 is higher in the later time. In this case, Wilcoxon test will be a better choice for the two group comparison.

Since the survival curves for the two groups differ primarily at longer survival times, the Wilcoxon test, which places more weight on shorter survival times, becomes less significant than the log-rank test. A likelihood ratio test which is assuming an exponential model is also included to compare the two group survivals. Figure 2 shows the graph of log survivor function estimates against survival time. Both curves resemble straight line through the origin, indicating the exponential model is possibly fit to the data. So The log likelihood test is appropriate to use in comparing the two groups. These three tests consistently suggest that the survival curves of the two groups are not significantly different. Notice that these tests are testing the entire survival curve over time, not for survival difference in certain time points.

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	0.1017	1	0.7498
Wilcoxon	0.0018	1	0.9664
-2Log(LR)	0.0634	1	0.8013

Table 3. Two groups survival comparison

VIEWTABLE: Work.Km_sur2 (Product-Limit Survival Estimates)

	drug	futime	Censoring Flag: 0=Failed 1=Censored	Survival Distribution Function Estimate	SDF Lower 95.00% Confidence Limit	SDF Upper 95.00% Confidence Limit	Stratum Number
1	1	0	.	1	1	1	1
2	1	41	0	0.9936708861	0.9559269615	0.9991060284	1
3	1	71	0	0.9873417722	0.9503388402	0.9968190992	1
4	1	131	0	0.9810126582	0.9422933105	0.9938364716	1
5	1	140	0	0.9746835443	0.9339580666	0.9904223569	1
6	1	179	0	0.9683544304	0.9256478964	0.9867049911	1
7	1	198	0	0.9620253165	0.9174307943	0.982758329	1
8	1	223	0	0.9556962025	0.9093188663	0.9786297285	1
9	1	334	0	0.9493670886	0.9013088563	0.9743517339	1
10	1	348	0	0.9430379747	0.893393328	0.9699478794	1
11	1	388	0	0.9367088608	0.8855641745	0.9654358536	1
12	1	400	0	0.9303797468	0.8778137511	0.9608293606	1
13	1	515	0	0.9240506329	0.8701351982	0.95613928	1
14	1	533	1	0.9240506329	.	.	1
15	1	673	0	0.9176778699	0.8624501514	0.9513487721	1
16	1	694	0	0.9113051069	0.8548288836	0.9464895757	1
17	1	732	1	0.9113051069	.	.	1

Display 2. SAS dataset for Kaplan Meier survival estimates by groups

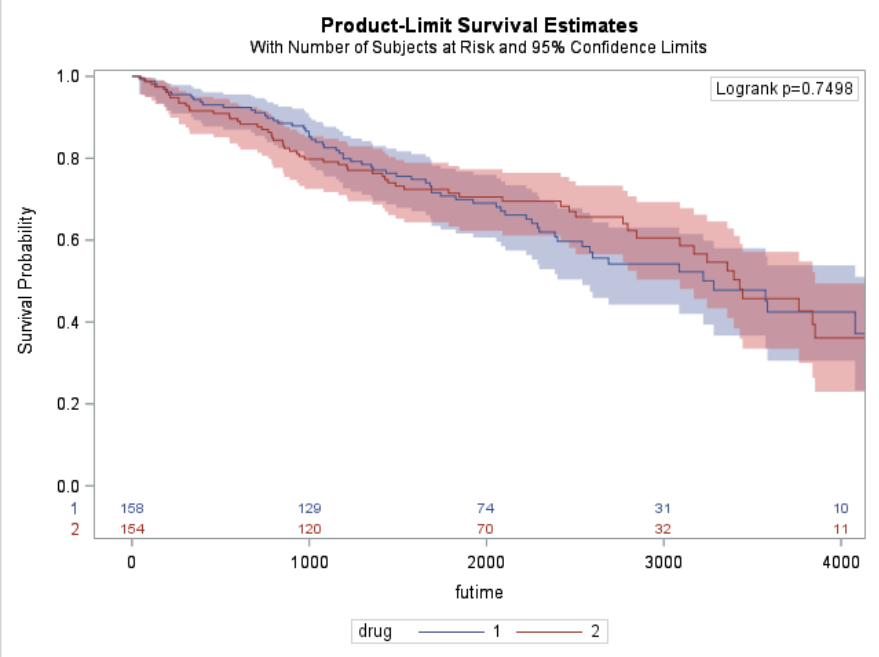

Figure 1. Plot of Estimated Survival Functions

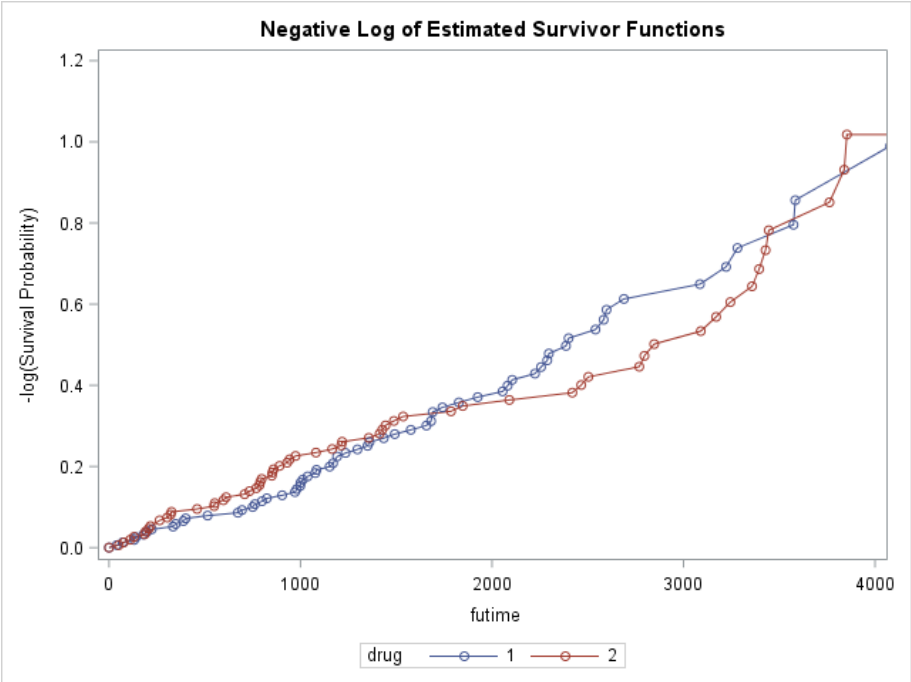

Figure 2. Plot of Estimated Negative Log Survivor Functions

PHREG PROCEDURE

The PHREG procedure perform survival regression based on Cox proportional hazard model (Cox 1972). Cox model is a semiparametric model and assumes the hazard of explanatory variables are constant over time.

The event time for each subject is defined by hazard function,

$$h_i\,(t) = h\,(t;\, z_i) = \; h_0\,(t)\exp(z_i'\beta)$$

Where h_0(t) is a unspecified baseline hazard function, z_i is vector of explanatory variables, β is the model coefficients.

2.1 How to assess individual risk factor on event risk

An example of commonly used PHREG codes are given below.

```
proc phreg data=pbc simple;
  where drug ~=.;
  class drug (ref='2') /param=ref;
  model futime*status(0,1)=drug;
  hazardratio drug/diff=ref cl=wald;
run;
```

Summary of the Number of Event and Censored Values			
Total	Event	Censored	Percent Censored
312	125	187	59.94

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
drug	1	1	0.05709	0.17916	0.1015	0.7500	1.059	drug 1

Hazard Ratios for drug			
Description	Point Estimate	95% Wald Confidence Limits	
drug 1 vs 2	1.059	0.745	1.504

Table 4. Output of PHREG procedure

2.2 How to select meaningful variable for risk adjustment.

Variable selection can be performed using SELECTION option in the MODEL statement. Stepwise selection uses alternative steps of forward and backward selection. You may specify the p value criteria for variables to enter model using option SLENTRY, and criterial for variables to stay in model using option SLSTAY. Some crucial variables can be forced to remain in model during selection process using option INCLUDE.

```
proc phreg data=pbc2 ;
  where drug ~=.;
  class drug (ref='Placebo') sex (ref='male') edema (ref='no')
ascites(ref='no') hepato(ref='no') spiders(ref='no') stage
(ref='4')/param=ref;
  model futime*status(0,1)= drug sex edema ascites hepato spiders stage age
logbili logalbumin logprotime platelet logsgot/ include=1 selection=stepwise
slentry=.15 slstay=.20 details;
  logbili=log(bili); logalbumin=log(albumin); logprotime=log(protime);
logsgot=log(sgot);
  format sex sex. drug drug. edema edema. ascites hepato spiders affirm.;

run;
```

Summary of Stepwise Selection						
Step	Effect	DF	Number	Score	Wald	Pr > ChiSq

	Entered	Removed		In	Chi-Square	Chi-Square	
1	logbili		1	2	155.9923		<.0001
2	logalbumin		1	3	33.3877		<.0001
3	age		1	4	17.6357		<.0001
4	logprotime		1	5	12.6612		0.0004
5	logsgot		1	6	3.7772		0.0520
6	edema		1	7	4.4604		0.0347
7	stage		3	8	5.4774		0.1400

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
drug	D-pencl	1	-0.11018	0.18790	0.3439	0.5576	0.896	drug D-pencl
edema	yes	1	0.47382	0.23042	4.2287	0.0397	1.606	edema yes
stage	1	1	-1.70138	1.02752	2.7417	0.0978	0.182	stage 1
stage	2	1	-0.46428	0.31712	2.1434	0.1432	0.629	stage 2
stage	3	1	-0.28418	0.22643	1.5751	0.2095	0.753	stage 3
age		1	0.03620	0.00929	15.1993	<.0001	1.037	
logbili		1	0.73800	0.11515	41.0736	<.0001	2.092	
logalbumin		1	-2.62254	0.76130	11.8667	0.0006	0.073	
logprotime		1	3.52896	1.19089	8.7811	0.0030	34.089	
logsgot		1	0.59723	0.25200	5.6167	0.0178	1.817	

Table 5. Output of variable selection in a Cox model

2.3 How to check model specification

In PHREG procedure, the ASSESS statement performs the graphical and numerical methods of Lin, Wei, and Ying [3] for checking the assumed Cox regression model. The methods show cumulative sums of martingale residuals over follow-up times or covariates. Supremum test is also provided to compare the simulation process given the assumed model to the observed process. If the observed process is within the patterns of the simulated paths and the p values from supremum test is not significant, it indicates that the model specification is valid.

2.3.1 Assess functional form of continuous variable in the model

To test if the function forms of continuous variables are correctly specified, option VAR can be used in the ASSESS statement.

```
assess var=(age logbili logalbumin logprotime  logsgot);
```

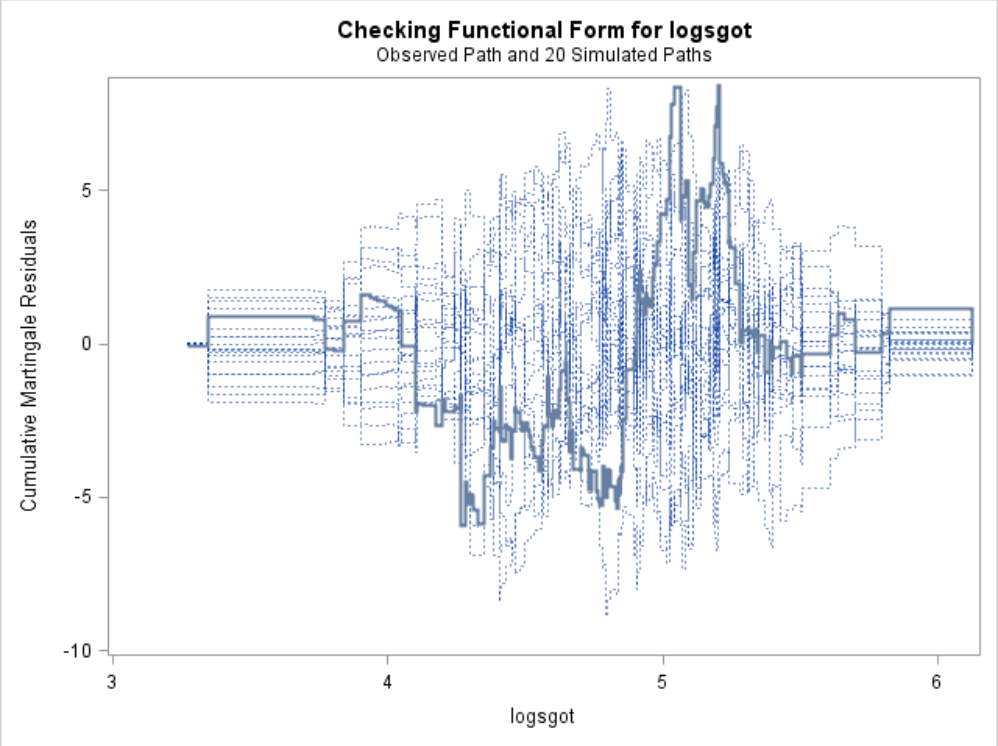

Figure 3. Checking functional form of a continuous variable in Cox model

2.3.2 Assess proportional hazard function

To test the proportional hazard assumption, option PH can be used in the ASSESS statement.

```
assess ph /resample=1000 seed=3538626 npath=20;
```

Supremum Test for Proportional Hazards Assumption				
Variable	Maximum Absolute Value	Replications	Seed	Pr > MaxAbsVal
drugD_pencl	1.1200	1000	3538626	0.1640
edemayes	1.5784	1000	3538626	0.0080
stage1	0.5833	1000	3538626	0.3300
stage2	1.0363	1000	3538626	0.2880
stage3	1.3854	1000	3538626	0.0760

Supremum Test for Proportional Hazards Assumption				
Variable	Maximum Absolute Value	Replications	Seed	Pr > MaxAbsVal
age	1.0090	1000	3538626	0.2150
logbili	1.2851	1000	3538626	0.1270
logalbumin	0.8449	1000	3538626	0.5210
logprotime	1.9162	1000	3538626	0.0030
logsgot	0.9886	1000	3538626	0.3380

Table 6. Supremum Test for Proportional Hazards Assumption

Use strata COX model to deal with variables with violating PH assumption. Stratified COX model assumes a different baseline hazard function for each stratum. For example, `strata edema;`

2.4 How to obtain predictive event risk over time - plot adjusted survival curve

For a patient with given condition, the survival probability over time can be calculated from the given model.

```
data var;
     format sex sex. drug drug. edema edema. ascites hepato spiders affirm.;
   input  drug  edema    stage age logbili logalbumin logprotime  logsgot
label;
   datalines;
  1 0 1 50 0.6 1.3 2.4 4.7  1
  0 0 1 50 0.6 1.3 2.4 4.7  0
  ;


proc phreg data=pbc2 plots (cl overlay )=survival atrisk;
  where drug ~=.;
  class drug (ref='Placebo')  edema (ref='no') stage (ref='4')/param=ref;
  model futime*status(0,1)=drug  edema stage age logbili logalbumin
logprotime  logsgot;
logbili=log(bili); logalbumin=log(albumin); logprotime=log(protime);
logsgot=log(sgot);
  format sex sex. drug drug. edema edema. ascites hepato spiders affirm.;
  *assess var=(age logbili logalbumin logprotime  logsgot);
  *assess ph /resample=1000 seed=3538626 npath=20;
  strata edema;
  baseline out=test1 covariates=var survival=a / group=drug rowid=label;run;
run;
```

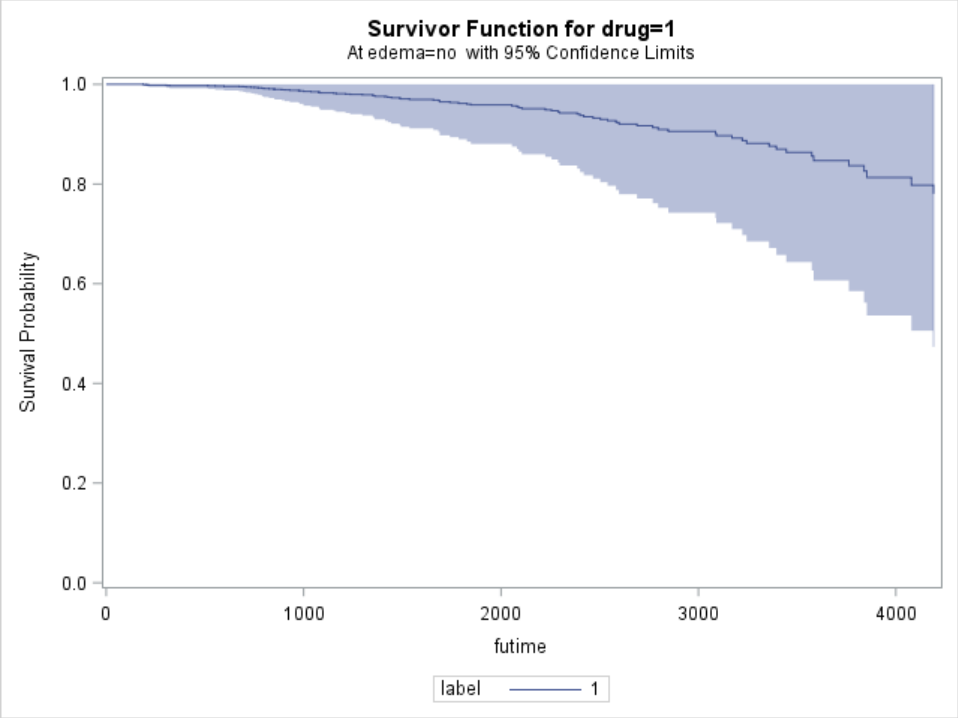

Figure 4. Plot predicted survival curve from a Cox model

2.5 Model time dependent variable

Time dependent variables can be included into the COX model in PHREG procedures. Two common ways for including such a time dependent variable in the model are 1) create a time dependent variable in the MODEL statement, this could be a status indicator change IF ELSE statement or a time interaction equation; 2) create a data set that includes the counting process in each time interval (start time, stop time).

In this example, although age is changing over time, age is treated as a time fixed covariate. As every subject is increasing age at a same pace, the hazard ratio of age at baseline time will be the same as the hazard ratios of age treating as a time dependent variable. We therefore use age at the baseline time as a time fixed covariate.

CONCLUSION

This paper uses a public available data to demonstrate use of SAS® LIFETEST and PHREG procedures for survival analysis. These SAS procedures allow users to perform various techniques in survival analysis, including Kaplan Meier method and Cox proportional hazard models.

REFERENCES

1. John P. Klein , M.L.M., *Survival Analysis: Techniques for Censored and Truncated Data.* 2005.
2. Allison, P.D., *Survival Analysis Using SAS: A Practical Guide.* Sas Inst., 2010.
3. Lin, D., Wei, L. J., and Ying, Z., *Checking the Cox Model with Cumulative Sums of Martingale-Based Residuals.* Biometrika, 1993. **80**: p. 557–572.

ACKNOWLEDGEMENTS

This paper acknowledges the support from Department of Cardiac Surgery in University of Michigan.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Xiaoting Wu (Ting), PhD, MS

Department of Cardiac Surgery

1500 E Medical Center Drive

Ann Arbor, MI 48109

734.936.7731

xiaotinw@med.umich.edu