Paper 75-2010

PROC LIFEREG or PROC PHREG

Dachao Liu, Northwestern University, Chicago, IL

ABSTRACT

Besides commonly used PROC LOGISTIC, PROC PROBIT, PROC GENMOD, PROC RELIABILITY and PROC LIFETEST, SAS® has PROC LIFEREG or PROC PHREG in doing survival analysis. They both contain REG, a reminder of regression analysis, and they both deal with time-to-event data. This sometimes makes us wonder when we should use PROC LIFEREG when we should use PROC PHREG, even for experienced statisticians who are using SAS. This paper will discuss this question by using some examples.

INTRODUCTION

The PROC LIFEREG and the PROC PHREG procedures both can do survival analysis using time-to-event data, what is the difference between the two. Before I get into the main topic, a little history about survival analysis may give us a clear picture of the development of survival analysis. Survival analysis can trace the origin of the development of actuarial science and demography back to the 17th century when the first life table was created by somebody. The Second World War not only witnessed initial military use of ROC analysis but also a new era of survival analysis, stimulated by interest in reliability (or failure time) of military equipment. 1958 saw the major break-through of survival analysis with the advent of the paper by Kaplan and Meier in which they proposed their famous product-limit estimator of the survival curve. Kaplan Meier method estimates and compares the survival function. In 1972 David Cox introduced the proportional hazards model to resolve covariates issue. Cox model can assess the effects of covariates on survival.

DEFINITIONS

Survival analysis has been used in many fields, hence acquiring many names: "event history analysis or duration models" (in social sciences like sociology, political science, and economics etc,) "reliability analysis or failure-time models," (engineering), "survival analysis(medical research)" Whatever the names, they all deal with duration of time of the variable of interest from an initial observation until the occurrence of a subsequent event. The variables of interest are marriage(in sociology), employment(in economics), machine brake-down(in engineering) and death(in medical research). This duration of time, or time interval between an initial point and subsequent event, often called failure, is known as the survival time. Survival analysis is to study this duration of time and the occurrence of event in time, the measurements of which are composed of survival data. Survival data record the lapsed time to some specific event. One frequently-encountered problem with these data is that the time to the event of interest is not always observable for all subjects. Or some units of observation are observed for variable lengths of time but do not experience the event (or endpoint) under study. These unobservable observations are called censored observations. Survival data is characterized by the censored observations. Censoring occurs when some study subjects cannot be observed due to one reason or another like some study subjects move away or die before the study is complete or refuse to participate any longer. These study subjects are said to be lost to follow-up. Censoring may also occur when time interval between an initial point and subsequent event is very long, the data may be analyzed before this subsequent event of interest has occurred in all study subjects. Censoring usually means the "end" of observation. Study subjects lost to follow-up(or the study ends) might have experienced a recurrence of the event at some time in the future, but the researcher would not know if or when this happened. Such observations are said to be right-censored. If study subjects' duration of time is known to be less than a certain duration, such observations are said to be left-censored. If study subjects' duration of time is known to be between an interval, such observations are said to be Interval-censored.

REGRESSION PART and MODEL PARAMETERS

SAS has PROC LIFEREG or PROC PHREG in survival analysis. They both contain REG, a reminder of regression analysis. In regression analysis, a response variable Y can be predicted by a linear function of a regressor variable X. We can estimate β_o, the intercept, and β_1, the slope, in

$$Y_i = \beta_o + \beta_1 X_i + \varepsilon_i$$

In SAS, this can be performed by

```
proc reg;
  model y=x;
run;
```

While the REG procedure is a general-purpose procedure for regression, the LIFEREG procedure and the PHREG procedure provide more specialized regression applications for survival analysis. Survival models can be usefully viewed as ordinary regression models in which the response variable is time. However, the censoring in survival data makes survival analysis different from regression analysis. Instead of using ordinary least squares (OLS) as in regression analysis, survival analysis is using the maximum likelihood function. Computing the likelihood function (needed for fitting parameters or making other kinds of inferences) is complicated by the censoring.

In the LIFEREG procedure, the model assumed for the response **Y** is

$$\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\sigma\varepsilon}$$

Where **Y** is a vector of response values, often the log of the failure times, **X** is a matrix of covariates or independent variables (usually including an intercept term), **β** is a vector of unknown regression parameters, **σ** is an unknown scale parameter, and **ε** is a vector of errors assumed to come from a known distribution such as the standard normal distribution.

The PHREG procedure performs regression analysis of survival data based on the Cox proportional hazards model. Cox's semiparametric model is widely used in the analysis of survival data to explain the effect of explanatory variables on hazard rates. The survival time of each member of a population is assumed to follow its own hazard function, $h_i(t)$ expressed as:

$$h_i(t) = h_o(t)exp(X_i'\boldsymbol{\beta})$$

where $h_i(t)$ is an arbitrary and unspecified baseline hazard function, X_i is the vector of explanatory variables for the ith individual, and **β** is the vector of unknown regression parameters that is associated with the explanatory variables. The vector **β** is assumed to be the same for all individuals. $X_i'\boldsymbol{\beta}$ is also known as the linear predictor of the form $\beta_1X_{1i} + \beta_2X_{2i} + \beta\kappa X\kappa_i$. Therefore, the above formula can also be written as:

$$ln[h_i(t) / h_o(t)] = \beta_1X_{1i} + \beta_2X_{2i} + \ldots + \beta\kappa X\kappa_i$$

To estimate **β**, Cox (1972) introduced the partial likelihood function, which eliminates the unknown baseline hazard $h_o(t)$ and accounts for censored survival times.

The partial likelihood of Cox also allows time-dependent explanatory variables. An explanatory variable is time-dependent if its value for any given individual can change over time. Time-dependent variables have many useful applications in survival analysis. We can use a time-dependent variable to model the effect of subjects changing treatment groups. Or we can include time-dependent variables such as blood pressure or blood chemistry measures that vary with time during the course of a study. We can also use time-dependent variables to test the validity of the proportional hazards model.

We can compare several groups and/or risk factors of study subjects in the PROC LIFETEST, but as the number of groups and/or factors gets bigger, it's extremely difficult to interpret the results. It's wise to resort to regression-like methods. That's the reason regression analysis has been integrated into survival analysis. In this sense, the PROC LIFETEST is to a one-way analysis of variance what the PROC LIFEREG is to two factor designs.

Since it's regression analysis, we can use FROWARD,BACKWARD and STEPWISE selection to select best variable for the model in PROC PHREG, a best tool for initial screening.

The baseline hazard portion of the model is nonparametric because no prior knowledge of its form is assumed. On the other hand the influences of the explanatory variables are described in a parametric linear-regression model with regression coefficients β_1 and β_2. The Cox model is hence said to be semi-parametric.

Survival data usually have some censored observations, otherwise we can use the REG procedure directly, because we will get the same coefficients and standard errors produced by the PROC REG and the PROC LIFEREG(a log-normal transformation is required)

DISCUSSION

Now let's discuss these two procedures used in survival analysis.

Prio to SAS version 6.10, there was no the PHREG procedure. The PHREG procedure came into being after the LIFEREG and was listed in the SAS documentation of SAS/STAT Software Changes and Enhancements in SAS version 6.11 in 1996.

Let's first compare statements in these two procedures up to SAS version9.22

Syntax: LIFEREG Procedure
PROC LIFEREG Statement
BAYES Statement
BY Statement
CLASS Statement
INSET Statement
MODEL Statement
OUTPUT Statement
PROBPLOT Statement
WEIGHT Statement

Syntax: PHREG Procedure
PROC PHREG Statement
ASSESS Statement
BASELINE Statement
BAYES Statement
BY Statement
CLASS Statement
CONTRAST Statement
EFFECT Statement
ESTIMATE Statement
FREQ Statement
HAZARDRATIO Statement
ID Statement
LSMEANS Statement
LSMESTIMATE Statement
MODEL Statement
OUTPUT Statement
Programming Statements
STRATA Statement
SLICE Statement
STORE Statement
TEST Statement
WEIGHT Statement

Words in italic are new statements added to SAS version 9.22.
At first glance, we see the PROC PHREG has more statements to use. Accordingly it can perform more functions than the LIFEREG procedure.

The LIFEREG procedure uses parametric method (maximum likelihood), dealing with data of left, right and interval censoring . It doesn't handle time-dependant covariates. It doesn't allow data step. It's easy to interpret the estimates of the survival function, because the explanatory variables in the model are fixed, not dependent on time. The PHREG procedure uses semi-parametric method (partial likelihood), dealing with data of only right censoring. It handles time-dependant covariates. It allows data step or the programming statements. It's not easy to interpret the estimates of the survival function, because the explanatory variables in the model are dependent on time. In 2010, SAS has come out the BPHREG procedure (Bayesian PHREG), which extends its stratification capability. Besides, bayesian analysis of survival models can be requested in both the LIFEREG and the PHREG procedures.

If there are no time-dependant covariates involved, we can use either the LIFEREG procedure or the PHREG procedure and the results would be very similar from each procedure. Here is an example:

Suppose we have a data set MIB, part of which is as follows:

TIMEM1	STATUS1	BONE_DX1	BONE_TX1	MIBG_DX1	MIBG_TX1
67.1869	0	12	2	15	1
32.7885	0	0	0	0	0
48.0657	0	4	0	4	0
40.7392	0	2	2	1	0
24.6407	1	12	2	0	0
43.7618	0	11	0	0	.
38.7023	1	17	10	19	17

We can use

```
proc lifereg data=MIB;
  model TIMEM1*STATUS1(0)= BONE_DX1 BONE_TX1 MIBG_DX1 MIBG_TX1;
 title"Overall Surval (1=Event 0=Censored)";
run;
```

(If the distribution is not specified, SAS will default to Weibull)

Or

```
proc phreg data=MIB;
  model TIMEM1*STATUS1(0)= BONE_DX1 BONE_TX1 MIBG_DX1 MIBG_TX1;
 title"Overall Surval (1=Event 0=Censored)";
run;
```

Here is the output from the LIFEREG procedure:

The LIFEREG Procedure

Analysis of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
BONE_TX1	1	-0.0164	0.2105	-0.4290	0.3962	0.01	0.9378
MIBG_DX1	1	0.1451	0.0889	-0.0291	0.3194	2.66	0.1026
MIBG_TX1	1	-0.1282	0.1252	-0.3736	0.1171	1.05	0.3057
Scale	1	0.4573	0.1512	0.2392	0.8743		
Weibull Shape	1	2.1867	0.7231	1.1437	4.1808		

Here is the output from the PHREG procedure:

The PHREG Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
BONE_DX1	1	0.29783	0.16174	3.3907	0.0656	1.347
BONE_TX1	1	-0.12263	0.53880	0.0518	0.8200	0.885
MIBG_DX1	1	-0.33924	0.19577	3.0027	0.0831	0.712

MIBG_TX1	1	0.35589	0.31346	1.2890	0.2562	1.427

The model information sections are almost the same in both procedures. They are not shown here.

In the last sections, Analysis of Maximum Likelihood Parameter Estimates, the PHREG procedure has no intercept estimate, because it uses partial likelihood estimation). The p-values are similar in both procedures, indicating that there are no significant effects of any variables. Unfortunately, we cannot compare the coefficients of the PHREG with those of the LIFEREG. Because the former is proportional hazard model while the latter is not.

Now when time-dependant covariates come into play, we can only use the PHREG procedure. Any variables can be covariates in a statistical model, for example age, gender, race, height, weight, education, income, treatment, blood pressures, etc.. If covariates change their values over time then they are called time-dependant covariates, such as age, income, treatment, blood pressures. Some covariates will not change their values overtime. They are called non-time-dependant covariates, such as gender and race. Covariates information is very important in predicating values of other variables. We can use gender, race, education or years of employment to predict income. With the non-time-dependant covariates, it's easy to establish a statistical model to make sound estimates. With the time-dependant covariates, a regular statistical model would not do the work. Cox developed a special model to deal with the time-dependant covariates. That's so called proportional-hazards model. After the LIFEREG procedure, SAS institute developed the PHREG procedure which handles time-dependant covariates. That's the main difference between the LIFEREG procedure and the PHREG procedure. The values of a time-dependant covariate are elusive, since they change over time. The PHREG procedure has programming statements, in other words, a data step within a procedure, to capture the instant value. This is another main difference between the LIFEREG procedure and the PHREG procedure.

Suppose we have a bone marrow transplant data BMT. (Data found on line at http://www.mcw.edu/biostatistics/Faculty/Faculty/JohnPKleinPhD/SurvivalAnalysisBook/DataSetsBothEditions.htm)
Graft type (1=allogenic, 2=autologous)
Disease type (1=Non Hodgkin lymphoma, 2=Hodgkins disease)
Time to death or relapse, days
Death/relapse indicator (0=alive, 1=dead)
Karnofsky score
Waiting time to transplant in months

We want to model the effects of subjects transferring from one treatment group to another by using a time-dependent variable. Bone marrow disease patients are waiting until a donor is available. Time spent on waiting is recorded as WTIME. A patient's status can be changed during the study from waiting for a transplant to being a transplant recipient, creating a time-dependent variable WSTATUS, which, in a mathematical sense, is the time-dependent covariate function Z(t). Z(t) takes the value of 0, if the patient has not received the transplant at time t and the value of 1, if the patient has received the transplant at time t. Since the values of WSTATUS changes over time, it's impossible to capture its values in a data step, therefore the variable WSTATUS cannot be created in the data step. But by using the programming statements in the PROC PHREG, it can be done as follows:

```
proc phreg data=BMT;
model TIME*STATUS(0)=WSTATUS KS;
if (TIME < WTIME) then WSTATUS=0;
else WSTATUS=1;
run;
```

Here is the output:

The PHREG Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
WSTATUS	1	0.07440	0.66356	0.0126	0.9107	1.077
KS	1	-0.05200	0.01149	20.4687	<.0001	0.949

from the output, we see transplantation appears to be associated with a slight increase in risk, but the effect is not significant (p=0.9107). KS adds significantly to the model (p<0.001). The risk decreases significantly with KS.

The Karnofsky Performance Scale Index classifies patients according to their functional impairment. The classification can be used to compare effectiveness of different therapies and to assess the prognosis in individual patients. The lower the Karnofsky score, the worse the survival for most serious illnesses.

It would be incomplete to discuss the PROC PHREG without mentioning the counting process style of input, also called counting process formulation, which is an important feature of the PROC PHREG. It involves a very complicated mathematical theory. Therefore this paper will not delve very deep into it. Normally survival analysis uses data of one record per subject; while on some special occasions, survival analysis uses data of more than one record per subject. This kind of data of one subject with multiple records will constitute the time-dependent repeated measurements.

Here is an example. Suppose we want to study the mortality of subjects in a situation, say an exposure to some toxins, we can record TSTART(Starting time), TSTOP(stop time), status(either death or censored) and x1, x2, x3, x4 and x5 (explanatory variables related to survival. We can use the PROC PHREG to accommodate this as follows:

```
proc phreg;
model (TSTART, TSTOP)*status(0) = x1 x2 x3 x4 x5;
run;
```

Here is another example. Suppose we want to study the effects of treatment on prostate cancer adjusting for PSA values. The study lasted 281 days. Each patient has his PSA measured from 1 to 5 times. The data is like this:

id	time	dead	treatment	p1	p2	p3	p4	p5
1	247	0	1	5.2	5.0	5.5	5.4	.
2	171	1	1	5.1	5.1	5.2	5.2	.
3	281	0	1	5.2	4.5	4.6	4.8	5.0
4	181	0	1	5.2	5.1	5.2	5.2	.
5	181	0	1	5.1	5.2	5.1	5.2	.
16	254	0	2	5.2	5.1	4.9	5.0	5.1
17	253	0	2	5.1	5.2	5.2	4.8	4.8
18	148	1	2	5.0	5.3	4.7	.	.
19	154	0	2	5.5	6.0	6.0	.	.
20	251	1	2	5.5	5.5	5.5	5.6	5.6
31	94	1	3	6.0	.	.	.	.
32	137	1	3	6.0	5.6	.	.	.
33	253	0	3	5.7	5.8	5.6	6.0	6.0
34	245	0	3	4.8	6.0	5.5	5.7	.
35	253	0	3	4.6	4.5	5.3	5.3	5.3

ID(Patient identification)
TIME(Survival time of the patient)
Dead(censoring status where 1=dead and 0=censored)
Treatment(radiation: 1= IMRT 2= EBRT and 3= 3DCRT)
(Intensity modulated radiation therapy (IMRT), electron beam radiation therapy (EBRT) and 3 dimensional conformal radiation therapy (3D-CRT) , newer versions of EBRT).
P1-P5(PSA values at 5 times that patients died) 5 deaths occurred at 94, 137, 148, 171, and 251 days.

After some data step, the data can be shaped like this:

id	time	dead	treatment	T1	T2	Status	PSA
1	247	0	1	0	94	0	5.2
1	247	0	1	94	137	0	5.0

1	247	0	1	137	148	0	5.5
1	247	0	1	148	247	0	5.4
2	171	1	1	0	137	0	5.1
2	171	1	1	137	171	1	5.2
3	281	0	1	0	94	0	5.2
3	281	0	1	94	137	0	4.5
3	281	0	1	137	148	0	4.6
3	281	0	1	148	171	0	4.8
3	281	0	1	171	281	0	5.0
4	181	0	1	0	94	0	5.2
4	181	0	1	94	137	0	5.1
4	181	0	1	137	181	0	5.2
5	181	0	1	0	94	0	5.1
5	181	0	1	94	137	0	5.2
5	181	0	1	137	148	0	5.1
5	181	0	1	148	181	0	5.2
16	254	0	2	0	94	0	5.2
16	254	0	2	94	137	0	5.1
16	254	0	2	137	148	0	4.9
16	254	0	2	148	171	0	5.0
16	254	0	2	171	254	0	5.1
17	253	0	2	0	94	0	5.1
17	253	0	2	94	148	0	5.2
17	253	0	2	148	253	0	4.8
18	148	1	2	0	94	0	5.0
18	148	1	2	94	137	0	5.3
18	148	1	2	137	148	1	4.7
19	154	0	2	0	94	0	5.5
19	154	0	2	94	154	0	6.0
20	251	1	2	0	148	0	5.5
20	251	1	2	148	251	1	5.6
31	94	1	3	0	94	1	6.0
32	137	1	3	0	94	0	6.0
32	137	1	3	94	137	1	5.6
33	253	0	3	0	94	0	5.7
33	253	0	3	94	137	0	5.8
33	253	0	3	137	148	0	5.6
33	253	0	3	148	253	0	6.0
34	245	0	3	0	94	0	4.8
34	245	0	3	94	137	0	6.0
34	245	0	3	137	148	0	5.5
34	245	0	3	148	245	0	5.7
35	253	0	3	0	94	0	4.6
35	253	0	3	94	137	0	4.5
35	253	0	3	137	253	0	5.3

Now the counting process style of input can be used

```
proc phreg data=psa1;
model (T1,T2)*Status(0)=Treatment  PSA;
run;
```

Here is the output:

The PHREG Procedure

Model Information

Data Set	WORK.PSA1
Dependent Variable	T1
Dependent Variable	T2
Censoring Variable	Status
Censoring Value(s)	0
Ties Handling	BRESLOW

Number of Observations Read	47
Number of Observations Used	47

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
47	5	42	89.36

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	24.203	23.448
AIC	24.203	27.448
SBC	24.203	26.667

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	0.7556	2	0.6854
Score	0.7892	2	0.6739
Wald	0.7641	2	0.6824

The PHREG Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
treatment	1	0.25074	0.66991	0.1401	0.7082	1.285
PSA	1	0.62356	1.26796	0.2419	0.6229	1.866

the effects of treatment on prostate cancer adjusting for PSA values is not statistically significant (p=0.7082).

Now a new procedure called the PROC SURVEYPHREG has been introduced. It is similar to the PHREG procedure because it also fits the proportional hazards model. It performs survival analysis survey data. But it doesn't have the feature of the counting process style of input yet.

CONCLUSION

The PROC LIFEREG and the PROC PHREG procedures both can do survival analysis using regression like method on time-to-event data to handle covariates. The PROC LIFEREG is the extension of the PROC LIFETEST (Kaplan Meier model); the PROC PHREG is regression analysis on Cox model with a piecewise exponential model as its update. Although it came into being later than the PROC LIFEREG, the PROC PHREG is more versatile. The PROC LIFEREG can only evaluate the effect of time-independent covariates and parametric fitting of survival time. The PROC PHREG can evaluate the effect of time-dependent covariates and semi-parametric fitting of survival time. The programming statements add versatility to the PROC PHREG so as to allow it to accommodate time-dependent repeated measurements of a covariate. Counting process formulation is a very important feature of the PROC PHREG procedure.

REFERENCES

Allison, Paul D. 1995. *Survival Analysis Using the SAS® System: A Practical Guide.* Cary, NC: SAS Institute Inc.

Cox, David.R. 1972. "Regression Models and Life Tables," *Journal of the Royal Statistical Society,* Series B (Methodological), 34: 187–220.

Kaplan, E. L. and Meier, P. 1958. "Nonparametric Estimation from Incomplete Observations," *Journal of American Statistical Association*, 53, 457–81.

SAS Institute Inc. 2010. "The LIFEREG Procedure" and "The PHREG Procedure". *SAS/STAT® 9.22 User's Guide, Second Edition*. Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

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

Dachao Liu
Northwestern University
Suite 1400
680 N Lake Shore Dr.
Chicago, IL 60611
Phone (312)503-2809
dachao-liu@northwestern.edu