PharmaSUG 2014 - Paper PO02

Survival Analysis Approaches and New Developments using SAS

Jane Lu, AstraZeneca Pharmaceuticals, Wilmington, DE
David Shen, Independent Consultant

ABSTRACT

A common feature of survival data is the presence of censoring and non-normality. It is inappropriate to analyze survival data by the conventional statistical methods such as linear regression or logistic regression, because of the characteristics of the survival data. When censoring, survival time can't be considered as a continuous variable. Linear regression, in which to compare mean time-to-event between groups, cannot deal with the influence from censored data correctly. Logistic method compares proportion of events between groups using odds ratio, but the differences in the timing of event occurrence are not considered. With unequal survival times, analyzing the probability of survival as a dichotomous variable by Chi-square test would fail to account for this non-comparability between subjects. This paper provides an overview of survival analysis and describes its principle and applications. Examples with SAS programming will illustrate the LIFEREG, LIFETEST, PHREG and QUANTLIFE procedures for survival analysis. The new developments including time-dependent covariates, recurrent events, quantile regression in identifying important prognostic factors for patient subpopulations and joint modeling of longitudinal such as quality of life and time-to-event data will also be discussed.

What is Survival Analysis

Survival analysis is used predominantly when the interest is in observing time to event. In biomedical sciences, the event of interest is often the time of death of an individual from the time of disease onset, diagnosis or time where a particular treatment (i.e. surgery, chemotherapy) was applied. Currently, the event has been a qualitative change at a particular time point. The event includes the time to a disease, time to disease-free status , time to onset of illness after exposure, time to recovery from illness, time to symptom relief, time to relapse (recurrence, progression), time to readmission, or transition above or below the clinical threshold of a meaningful continuous variable (i.e. CD4 counts, plate counts after marrow transplantation). In clinical trials, the starting reference time includes randomization, first treatment, onset of exposure, diagnosis or surgery.

What Does Survival Analysis Do

When study period is long enough to observe the survival time of all subjects, we may use more common methods such as t-test or regression analysis by considering survival time as a continuous variable.
Survival analysis can

1. Estimate time-to-event for a group of individuals
2. Compare time-to-event between two or more groups
3. Assess the relationship of covariates to time-to-event.

Functions Describing Survival Distributions

1. Cumulative Distribution Function (cdf)

The cdf is defined as F(t) = P (T < t). It describes the probability that the random variable T (time of death) will be less than or equal to some time that we choose. F(t) is a non-decreasing function of t, and as t approaches ∞, F(t) approaches 1. F(t) has the probability values between 0 and 1.

2. Probability Density Function (pdf)

The probability of the failure time occurring at exact time t (out of the whole range of possible t's).

$$f(t) = \lim_{\Delta t \longrightarrow 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}$$

The pdf is the derivative of the cdf, f(t) = d F (t) / dt. It is very useful in describing the continuous probability distribution of a random variable. The probability P(a < T < b) is the area under the curve between time a and time b.

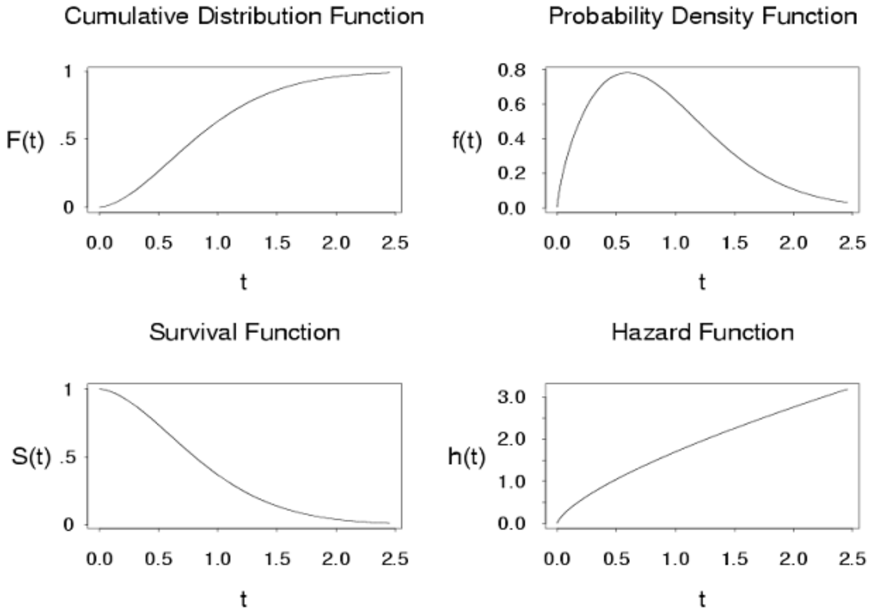


Figure 1. Survival Distribution Functions

3. Survival Function (sdf)

In survival analysis, survival function is of the most interest, and it which is defined as S(t) = P(T > t). The survival function is the probability that the time of death is later than some specified time. S(t) is positive and in the range from 1 to 0. S(0) = 1 and as t approaches ∞, S(t) approaches 0. S(t) = P(T > t) = 1 - F(t)

4. Hazard Function

S(t) is the prevalence of not failing, while h(t) is the incidence of failure. The hazard rate is an un-observed variable. It is the fundamental dependent variable controling both the occurrence and the timing of the events. The hazard function is the conditional probability of failure in the next interval given survival to start of that interval. Models for survival analysis can be built from a hazard function, which measures the risk of failure of an individual at time. The hazard function h(t) is given by the following formula:

$$h(t) = \lim_{\Delta t \longrightarrow 0} \frac{P(t \le T < t + \Delta t / T \ge t)}{\Delta t}$$

The hazard function seems to be more intuitive to use in survival analysis than the pdf because it attempts to quantify the instantaneous risk that an event will take place at time t given that the subject survived to time t. The hazard function is always positive and when it is zero, it implies failure is impossible at that time. It can be constant over time (Exponential distribution), increasing or deceasing with time (Weibull distribution). Higher values of h(t) carry an increasing risk of failure. The survival function and the hazard function are closely connected. Larger values of h(t) will yield lower values of S(t) since S(t) measures the risk of not failing. The hazard is the basis for regression modeling. It is related to other functions. h(t) = P(t < T < (t + Δ) | T >t) = f(t) / (1 - F(t)) = f(t) / S(t). Once we have modeled the hazard function, we can easily obtain the other functions.

Nature of Survival Time Data

Survival data measures lifetime or the length of time until the occurrence of an event. They are usually not normally distributed and also involve censoring. A censored observation is defined as an observation with incomplete information. Survival time is often censored. Survival time can be greater than a certain amount (right censored), less than a certain amount (left censored), or within a certain range (double censored). Of the three types of censoring methods, right censoring is the most common. Right censoring occurs because subjects are removed before failure or because failure occurs after the end of data collection. The purpose of survival analysis is to follow subjects over time and observe at which timepoint they experience the event of interest. Often times, the study is not long enough for an event to occur for all subjects, due to a number of reasons. Subjects may drop out of the study for reasons unrelated to the study (i.e. patients moving to another area and leaving no forwarding address). If the subject were able to stay in the study then it would have been possible to observe the event eventually. The censored survival time is usually indicated by the following censoring variable. For example, the variable LIFETIME represents either a failure time or a censoring time. The variable CENSOR is equal to 0 if the value of LIFETIME is a failure time, and it is usually equal to 1 if the value is a censoring time.

Another characteristic of survival data is that the response cannot be negative. This suggests that a transformation of the survival time such as a log transformation may be necessary or that special methods may be more appropriate than those that assume a normal distribution for the error term. It is especially important to check any underlying assumptions as part of the analysis because some of the models used are very sensitive to these assumptions.

Considerations of Survival Analysis Methods

Special considerations need to be taken when analyzing survival time data. Censoring and non-normality, cause great difficulty when trying to analyze survival data using traditional statistical models such as multiple linear regression. Data that have censored or truncated observations are said to be incomplete, and the analysis of such data requires special techniques.

Data with censored observations cannot be analyzed by ignoring the censored observations because, among other considerations, the longer-lived individuals are generally more likely to be censored. Logistic regression analysis could be applied to quantify the importance of certain covariates in classifying individuals into groups, those that did or did not experience the event during the period of observation. However, this approach can result in considerable loss of information because differences in the timing of event occurrence are not considered. Alternatively, one could use linear regression analysis to identify covariates that influence survival times. The major drawback is that survival data are often censored, i.e., they contain observations for which one does not know when the event occurrs. With conventional statistical methodology, censored observations would either have to be deleted, or one would set their survival times to the total time period from onset to termination of the study.

The proper method of survival data analysis is to take censoring into account and correctly use censored observations as well as uncensored observations. The likelihood-based parameter estimation methods used in survival analysis can effectively extract relevant information from both censored and uncensored observations, thereby producing reliable parameter estimates. Survival analysis is also the only method that can readily accommodate time-dependent covariates, i.e., independent variables whose values change during the course of the study. Disease severity is a time-dependent covariate, given that severity values change over time and they do so differently for each individual.

There are basically three methods to analyze survival data, namely parametric, non-parametric and semi-parametric. Parametric methods assume the knowledge of the distributions of the survival times e.g. Exponential, Weibull, Normal, Log-logistic and Gamma. Non-parametric models make no assumptions of the distribution of the survival time e.g. the Kaplan and Meier estimators. Semi-parametric models assume a parametric form for the effects of the explanatory variables but make no assumptions of the distributions of the survival time.

Survival Analysis Procedures

There are three SAS procedures for analyzing survival data: LIFEREG, LIFETEST and PHREG. PROC LIFETEST is a nonparametric procedure for estimating the distribution of survival time, comparing survival curves from different groups, and testing the association of survival time with other variables. PROC LIFEREG and PROC PHREG are regression procedures for modeling the distribution of survival time with a set of concomitant variables.

Table 1. SAS Procedures for Survival Analysis Applications

	PROC LIFEREG	PROC LIFETEST	PROC PHREG
Assumption of underlying survival time distribution	Must be specified (e.g., exponential, Weibull, gamma)	Shape not specified	Shape not specified
Model formulation	$\log_e T = \beta_0 + \beta_1 X_1 + \beta$		$\log_e h(t) = \log_e \lambda_0(t) + \beta_1 X_1 + \beta_2 X_2$
Estimation method	Maximum likelihood (parametric)	Kaplan-Meier or Life table method	Partial likelihood (semi-parametric)
Modeling the effect of discrete and/or continuous covariates on survival times. Hazard ratio, parameter estimates and associated significance levels	Ready		Ready
Effect of covariates	Multiplicative effect on survival times		Multiplicative effect on hazard functions
Time-dependent covariates	Not available		Readily included
Estimation plots of survival distribution		s(t), h(t); and derived median residual lifetime	
Test statistics for hypothesis of equality among groups		Logrank, Wilcoxon, and likelihood ratio tests.	

The LIFEREG Procedure – Parametric Model

Survival data consists of a response variable that measures the event time and possibly a set of independent variables thought to be associated with the failure time variable. These independent variables (covariates, or prognostic factors) can be either discrete, such as gender or race, or continuous, such as age or temperature. The purpose of survival analysis is to model the underlying distribution of the failure time variable and to assess the dependence of the failure time variable on the independent variables.

For the censored observations, an additional variable is incorporated into the analysis indicating which responses are observed event time and which are censored time.

The LIFEREG procedure fits parametric models to failure time that can be right-, left-, or interval-censored. The models for the response variable consist of a linear effect composed of the covariates and a random disturbance term. The baseline distribution of the error term can be specified as one of several possible distributions, including, but not limited to, exponential, Weibull, normal, and logistic distributions.
Some relations among the distributions are as follows:
The gamma with Shape=1 is a Weibull distribution.
The gamma with Shape=0 is a lognormal distribution.
The Weibull with Scale=1 is an exponential distribution.

The following artificial data are for a study of survival time. Subjects were randomized into two treatments (treat 1 = placebo, treat 2=medication). The observation was right-censored if death was not recorded.
Before beginning any modeling, preliminary investigations, such as, univariate analysis, should be done with graphics or other descriptive methods. We also used OUTPUT statement to create a new dataset containing hazard estimates and checked the hazard rate estimates with time.

```
title 'Lifereg with WEIBULL distribution';
proc lifereg data= surv ;
  model time*censor(0)= /dist=weibull;
  output out= hazard;
run;

proc sort data =hazard;
```

```
  by treat time;
run;

proc gplot data = hazard;
  plot h* time =treat;
  symbol1 i=sm50 line=1 c =red ;
  symbol2 i=sm50 line=2 c =blue;
run; quit;
```

The approximately constant hazard rates over time in the h(t) plot (see Figure 2(d)) suggested the time distribution may be modeled as exponential.

```
title 'Lifereg with EXPONENTIAL distribution';
proc lifereg data= surv ;
  class treat sex ;
  model time*censor(0)= age treat | sex /dist=exp;
run;
```

By default , PROC LIFEREG uses Weibull distribution. If the distribution is set to exponential (i.e. assume scale=1), the Lagrange test will be automatically conducted to test whether exponential model is adequate relative to the Weibull model.

```
          Lagrange Multiplier Statistics
     Parameter     Chi-Square    Pr > ChiSq
      Scale            0.5665        0.4516
```

CLASS statement determines which independent variables are treated as categorical, in order to test their interactions if any. In this model, variable treat and sex are defined as categorical, treat | sex is equivalent to treat sex treat*sex.

```
             Type III Analysis of Effect
                                  Wald
      Effect       DF    Chi-Square    Pr > ChiSq
      age           1       46.0552        <.0001
      treat         1       14.6338        0.0001
      sex           1        2.3947        0.1217
      treat*sex     1        0.0467        0.8290
```

Since treat*sex interaction is not significant, it is removed from the model and we get the following model parameters:

```
                     Analysis of Parameter Estimates
                           Standard   95% Confidence     Chi-
Parameter     DF Estimate     Error       Limits       Square Pr > ChiSq

Intercept      1   1.7601   0.3757   1.0236   2.4965   21.94     <.0001
age            1  -0.0481   0.0072  -0.0622  -0.0339   44.31     <.0001
treat          1   0.2875   0.0747   0.1411   0.4339   14.81     0.0001
sex            1  -0.1239   0.0746  -0.2702   0.0224    2.76     0.0969
Scale          0   1.0000   0.0000   1.0000   1.0000
Weibull Shape  0   1.0000   0.0000   1.0000   1.0000
```

If scale parameter is greater than 1, hazard decreases with time. Shape parameter is just 1/scale parameter. Scale of 1.0 in our example makes a Weibull an exponential. Weibull (and thus exponential) are proportional hazards models, so hazard ratio can be calculated. For other parametric models, you cannot calculate hazard ratio (hazards are not necessarily proportional over time).

Hazard Ratio (treated vs. control): HR 2:1 = e-β = e-0.2875 = 0.75
Risk Reduction: RR 2:1 = (1 - HR) *100 = 25%.
Interpretation: median time to death was decreased 25% in treated group or equivalently, mortality rate is 25% lower in treated group, compared to placebo control group.

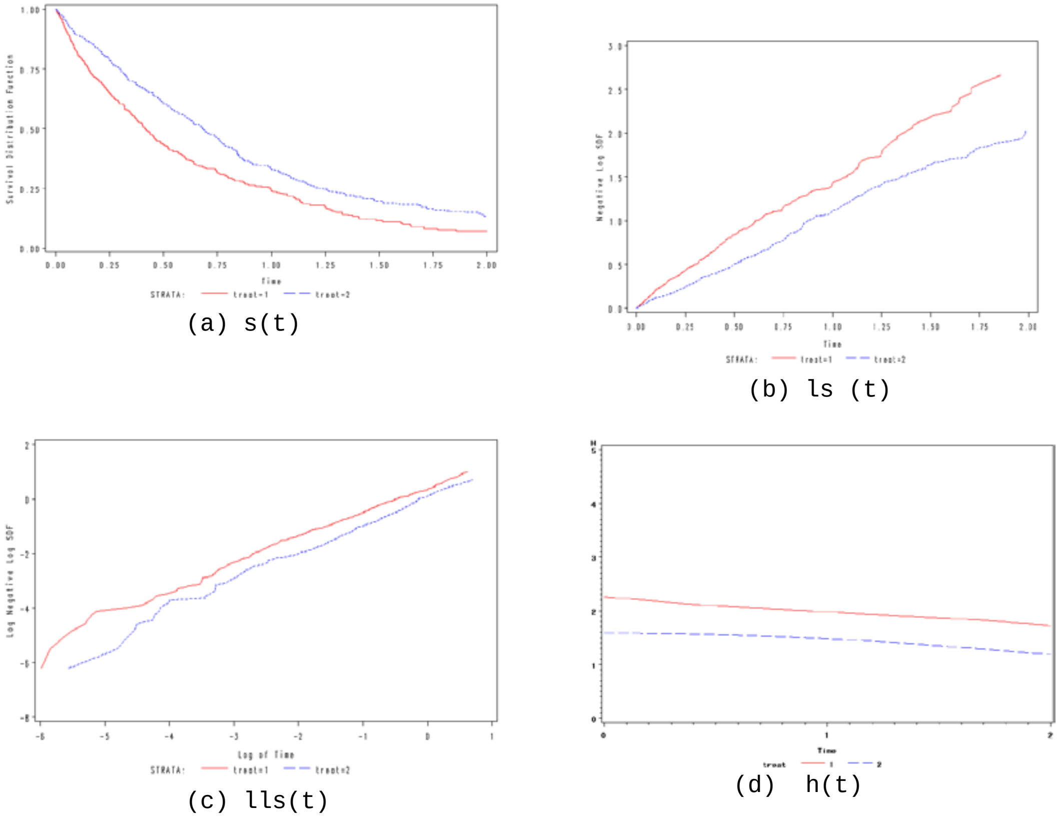
The LIFETEST Procedure – Nonparametric Model

Usually, the first step in analyzing survival data is to estimate the distribution of the survival time, S(t) = P(T > t). The LIFETEST procedure is used to compute nonparametric estimates of the survivor function either by Kaplan-Meier method (also known as the product-limit method) or by the life-table method (also called the actuarial method).

1. PROC LIFETEST can directly generate estimated survival plots. The PLOTS= option requests the function against time (by specifying S), a plot of the negative log of the estimated survivor function against time (by specifying LS), and a plot of the log of the negative log of the estimated survivor function against log time (by specifying LLS). The LS and LLS plots provide an empirical check of the appropriateness of the exponential model and the Weibull model. If the exponential model is appropriate, the LS curve should be approximately linear through the origin. If the Weibull model is appropriate, the LLS curve should be approximately linear. If there is more than one stratum, the LLS plot may also be used to check the proportional hazards model assumption. Under this assumption, the LLS curves should be approximately parallel across strata. If the life-table method is chosen, the estimates of the probability density function and the hazard function can also be computed. The generation of Kaplan-Meier and life table curves is primary for univariate analysis of the timing of events.
2. An important task in the analysis of survival data is the comparison of survival curves. It is of interest to determine whether the underlying populations of k (k≥2) samples have identical survivor functions. PROC LIFETEST provides non-parametic k-sample tests based on weighted comparisons of the estimated hazard rate of the individual population under the null and alternative hypotheses. PROC LIFETEST provides log-rank test, Wilcoxon test. Stratified tests can be specified to adjust for prognostic factors that affect the event rates in different populations. A likelihood ratio test, -2log(LR), based on an underlying exponential model, is also included to compare the survival curves of the samples.
3. The covariates may be related to the failure time. PROC LIFETEST can test the association between the covariates and the lifetime variable, by computing two such test statistics: censored data linear rank statistics based on the exponential scores and the Wilcoxon scores. The corresponding tests are known as the log-rank test and the Wilcoxon test, respectively. These tests are computed by pooling over any defined strata, thus adjusting for the stratum variable.

The null hypothesis of the logrank test is Ho: S1(t) = S2(t) for all values of t. The null hypothesis assumes that two curves should overlay exactly. The logrank test does not require that the groups have so-called proportional hazards.

```
proc lifetest data=surv  plots=(s,ls,lls)  censoredsymbol=none;
  time time*censor(0);
  strata treat;
run;
```



(a) s(t)

(b) ls (t)

(c) lls(t)

(d) h(t)

Figure 2. Survival Time Distributions

Both cumulative hazard functions appear mostly linear. This indicates relatively constant hazards for both treatment groups, with a larger hazard for treatment group 1. Parallel lines indicate that hazards between two groups are proportional over time, thus the hazard ratio can be calculated (necessary assumption for calculation of Hazard Ratios).

Log-rank test is basically a Cochran-Mantel-Haenszel chi-square test, it is more appropriate when long term effects are anticipated. Log-rank test has most power to test differences that fit the proportional hazards model—so works well as a set-up for subsequent Cox regression. Wilcoxon is just a version of the log-rank test that weights strata by their size (giving more weight to earlier time points). Wilcoxon is more sensitive to differences at earlier time points, so it is the preferred test for picking up short-term differences. Likelihood Ratio test only fits under the assumption of exponential distribution (constant hazard). Results of linear rank tests are shown below. The treatment effect is statistically significant for both Wilcoxon test (p=0.0021) and log-rank test (p=0.0030).

```
 Testing Homogeneity of Survival Curves for Time over Strata
                 Test of Equality over Strata
          Test      Chi-Square      DF    Chi-Square
          Log-Rank      9.4428       1      0.0021
          Wilcoxon      8.8257       1      0.0030
          -2Log(LR)     4.6693       1      0.0307
```

The TEST statement can specify a list of numeric (continuous) covariates that you want tested for association with the failure time.

```
proc lifetest data=surv  ;
  time time*censor(0);
  test age ;
run;
```

Two sets of rank statistics are computed here. Age is not statistically significant for both Wilcoxon test (p=0.0764) and log-rank test (p=0.0852).

```
Rank Tests for the Association of Time with Covariates
         Univariate Chi-Squares for the Wilcoxon Test

               Test       Standard                     Pr >
Variable    Statistic    Deviation    Chi-Square    Chi-Square
  age         -108.0       60.9691       3.1399       0.0764

         Univariate Chi-Squares for the Log-Rank Test

               Test       Standard                     Pr >
Variable    Statistic    Deviation    Chi-Square    Chi-Square
  age         -136.2       79.1465       2.9629       0.0852
```

The PHREG Procedure – Semi-parametric

$$h_i(t) = \lambda_0(t) e^{\beta_1 x_{i1} + \ldots + \beta_k x_{ik}}$$

The PHREG procedure fits Cox proportional hazards model to survival data with right censoring. The model is semi-parametric since it does not require any assumption of survival time distribution but it involves a finite number of regression parameters. It does not choose any particular probability models to represent survival time, and is therefore more robust than parametric method. All Cox regression requires is an assumption that ratio of hazards is constant over time across groups. It assumes the hazard functions of two groups are parallel over time. The ratio of hazard rates is called hazard ratio or relative risk. We don’t need to know anything about overall shape of risk/hazard over time, but the proportionality assumption can be sometimes restrictive. Only when hazard functions are parallel, the model can produce covariate-adjusted hazard ratios. Cox models can accommodate both discrete and continuous measures of event time the effect of covariates on the hazard rate but leaves the baseline hazard rate unspecified. It estimates relative rather than absolute risk without knowledge of absolute risk. Cox model uses the method of maximum partial likelihood to estimate the parameters so that we can estimate the coefficients without having to specify the baseline hazard function. The PL assumes no tied values among observed survival time, but it

$$HR_{i,j} = \frac{h_i(t)}{h_j(t)} = \frac{\lambda_0(t) e^{\beta_1 x_{i1} + \ldots + \beta_k x_{ik}}}{\lambda_0(t) e^{\beta_1 x_{j1} + \ldots + \beta_k x_{jk}}} = e^{\beta_1 (x_{i1} - x_{j1}) + \ldots + \beta_1 (x_{ik} - x_{jk})}$$

is not often the case with real data. SAS option on the model statement provides four methods: ties=exact/efron/breslow/discrete. EXACT computes the exact conditional probability under the proportional hazard assumption that all tied event time occurs before censored time of the same value or before larger values. This method may take a considerable amount of computer resources. If ties are not extensive, the EFRON and BRESLOW methods provide satisfactory approximations to the EXACT method for the continuous time-scale model. Breslow is SAS default, Breslow does not do well when the number of ties at a particular time point is a large proportion of the number of cases at risk. Efron's approximation gives results that are much closer to the EXACT method results than Breslow's, so Efron is preferred over Breslow. If the time scale is genuinely discrete, you should use the DISCRETE method. If there are no ties, all four methods result in the same likelihood and yield identical estimates.

```
proc phreg data =surv ;    /*First procedure*/
  model time * censor (0) = sex treat sextreat /ties = breslow;
    sextreat=sex*treat ;
run;

proc phreg data =surv ;             /*Second procedure*/
  model time*censor(0) = age sex treat / ties=Efron;
  output   out=a survival=s logsurv=ls loglogs=lls ;
  baseline out=b survival=s logsurv=ls loglogs=lls ;
run ;
```

The first procedure checks the interaction between treat and sex, the p-value of 0.55 indicates the interaction is not statistically significant. Let’s look at the output from the second procedure:

```
                                Testing Global Null Hypothesis: BETA=0

                         Test                 Chi-Square       DF     Pr > ChiSq

                         Likelihood Ratio        60.1548        3         <.0001
                         Score                   60.8132        3         <.0001
                         Wald                    60.9183        3         <.0001

                                Analysis of Maximum Likelihood Estimates

                      Parameter    Standard                              Hazard   95% Hazard Ratio
   Variable   DF       Estimate       Error  Chi-Square  Pr > ChiSq       Ratio   Confidence Limits

   age         1        0.05040     0.00740     46.4167      <.0001       1.052     1.037     1.067
   sex         1        0.12770     0.07472      2.9210      0.0874       1.136     0.981     1.315
   treat       1       -0.29932     0.07521     15.8371      <.0001       0.741     0.640     0.859
```

The information on Testing Global Null Hypothesis (H0: ß1 = … = ßp = 0) gives the result of the three tests, (partial) Likelihood ratio test, Score test and Wald test. All three tests lead to very small p-values. We can conclude that at least one of the coefficients is not 0. □ The main part of this ou estimated standard errors, chi-square statistics and p-values of Wald test (H0: ßj =0). For example, the estimated hazard ratio for TREAT is 0.741, which means that the hazard of death for those who received treatment is about 74% of the hazard for those who took placebo.

The RL option provides for each explanatory variable, 95% confidence limits for the hazards ratio ($e^{\hat{\beta}_i}$).The OUTPUT statement creates a new SAS data set containing statistics calculated for each observation. These can include the estimated survival distribution estimates, linear predictor and its standard error, residuals, and influential statistics.

The validity of the OH model results depends on non-violation of this assumption. This thus calls for testing for non-violation of the assumption in any analysis that we do. The test for non-violation of PH assumption can either be done formally or graphically. The formal test is done by interacting each covariate of interest with the log(time) variable. Using log(time) instead of time variable ensures that there is no numerical overflow. A non-significant coefficient in the interaction covariate means that the PH assumption is satisfied.

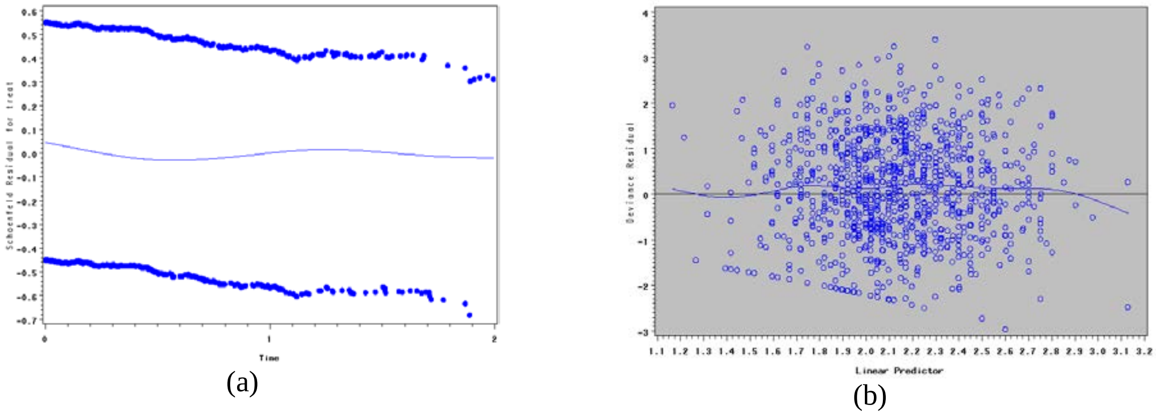
It is recommended to also graphically examine the residuals, which can be done using the Schoenfeld and/or weighted Schoenfeld residuals. The weighted Schoenfeld residuals method is preferred if there are tied observations. This can be done by examining the correlation coefficients. RESSCH = in SAS provides Schoenfeld residuals. These residuals are useful in assessing the proportional hazard assumption. WTRESSCH is the weighted Schoenfeld residuals. These residuals are useful in investigating the nature of non-proportionality if the proportional hazard assumption does not hold.

It is also good to formally carry out a linear-trend test on the residuals to investigate the lack of fit of a model. A non-significant P-value for the correlation coefficients means there is no trend. You can obtain martingale and deviance residuals for the Cox proportional hazards regression analysis by requesting that they be included in the OUTPUT data set. You can plot these statistics and look for outliers.

```
proc phreg data =surv;
  model time*censor(0) = age treat sex ;
  output out=Outs ressch=schage schtreat schsex
                  resdev=dev resmart=mart xbeta=xb;
run;

title “Schoenfeld residuals”;
axis1 label=(angle=90);
proc gplot data=Outs;
plot (schage schtreat schsex) * time  / vaxis=axis1  ;
symbol1 value=dot h=1 i=sm60S;
run;

title “Residual analysis”;
proc gplot data=Outp;
  plot (mart dev)*xb / vref=0 cframe=ligr;
  symbol1 value=circle c=blue;
run;
```



(a)

(b)

Figure 3. Treatment Schoenfeld Plot & Residual Plot

PLOT Schoenfeld residuals against time with a smooth line (cubic spline) fit to the points. Smoothing line helps to visualize the plots. There is no clear pattern with time. The plot of residual against predictor scores shows there is no indication of a lack of fit of the model to individual observations.

Considerations in Model Building

In any data analysis it is always a great idea to do some univariate analysis before proceeding to more complicated models. In survival analysis it is highly recommended to look at the Kaplan-Meier curves for all the categorical predictors. This will provide insight into the shape of the survival function for each group and give an idea of whether or not the groups are proportional (i.e. the survival functions are approximately parallel). We also consider tests of equality across strata to explore whether or not to include the predictor in the final model. For categorical variables we will use log-rank test of equality across strata which is a non-parametric test. For continuous variables we will use a univariate Cox proportional hazard regression which is a semi-parametric model. We will consider including the predictor if the test has a p-value of 0.2 - 0.25 or less. We are using this elimination scheme because all the predictors in the data set are variables that could be relevant to the model. If a predictor has a p-value greater than 0.25 in a univariate analysis, it is highly unlikely that it will contribute anything to a model which includes other predictors.

If the proportional hazard assumption in Cox regression model is valid, the significant predictors for survival can be identified. Five variable selection methods are available in PROC PHREG. When SELECTION=FORWARD, PROC PHREG first estimates variables computes the adjusted chi-square statistics for each variable not in the model and examines the largest of these statistics. If it is significant at the SLENTRY= level, the corresponding variable is added to the model. Once a variable is entered in the model, it is never removed from the model. The process is repeated until none of the remaining variables meet the specified entry level. When SELECTION=BACKWARD, parameters for the complete model as specified in the MODEL statement are estimated. Results of the Wald test for individual parameters are examined. The least significant variable that does not meet the SLSTAY= level for staying in the model is removed. Once a variable is removed from the model, it remains excluded. The process is repeated until no other variable in the model meets the specified level for removal. The SELECTION=STEPWISE option is similar to the SELECTION=FORWARD option except that variables already in the model do not necessarily remain. Variables are entered into and removed from the model in such a way that each forward selection step can be followed by one or more backward elimination steps. The stepwise selection process terminates if no further variable can be added to the model or if the variable just entered into the model is the only variable removed in the subsequent backward elimination.

The following is a sample SAS code of modeling with stepwise method to include or remove the covariates at each step. The variables that are statistically significant in the final step of the selection procedure are the significant covariates in the model.

```
proc phreg  data= surv;
  model time*censor(0)= var1 var2 … varn / selection = stepwise slentry = 0.25
                                           slstay = 0.15 details ;
run;
```

After determining the covariates in the final model, PROC PHREG can be run with the OUTPUT and BASELINE statements. One may be interested in computing the survival function estimates for a set of covariates with specific values other than the mean, or may want to use the established Cox regression model to generate predicted survival curves for subjects not in the study. The BASELINE statement can obtain the survivor function for such a set of explanatory variable values. The various sets of explanatory variable values should be contained in a SAS data set, and loaded with the COVARIATES= data set name. By default the data set contains the survival function estimates corresponding to the means of the covariates for each stratum. The survival estimates then can be displayed by PROC GPLOT for comparison.

New Developments

1). QUANTLIFE Procedure – Examine potential heterogeneous effects

Quantile regression provides a flexible way to capture heterogeneous effects in the sense that the tails and the central location of the conditional distributions can vary differently with the covariates. Thus, Quantile regression offers a powerful tool in survival analysis, where the lifetimes are skewed, or extreme survival time is suspected to be related with the covariates of interest.

The classical quantile regression method is not appropriate when the observations are incomplete, as is the case with censored data in survival analysis. The QUANTLIFE procedure implements appropriate quantile regression methods with reproducibility of resampling method that is used for statistical inference, to model the relationship between the response and the predictors.

The following statements of PROC QUANTLIFE can explore the relationship between survival time and two covariates of treatment and age at different quantiles (25th, 50th, and 75th percentiles).

```
proc quantlife data=surv plots=quantplot seed=12345;
  model time*censor(0)= treat age / quantile=(0.25 0.5 0.75);
  Effect: test treat;
run;
```

2). Crossover treatment – Adjust for crossover bias

In RCTs patients are allowed to switch from the control treatment to the new intervention after a certain timepoint (eg disease progression) due to ethical and other reasons. Crossover will cause the overall survival (OS) estimates confounded, and is likely to result in an underestimate of the treatment effect. The statistical methods to adjust for selective crossover in survival analysis may include

- Exclusion of patients with crossover treatment
- Censoring of patients at crossover treatment
- Inclusion of crossover treatment as a time-dependent covariate in Cox model
- Inverse Probability of Censoring Weights (IPCW)

- Rank Preserving Structural Failure Time Models (RPSFTM)
- Structural Nested Models (SNM)

The last three are more complex statistical methods. The following statements fit the third method for crossover as the time-dependent covariate in Cox regression analysis:

```
proc phreg data= Surv;
  class Treat;
  model Time*Censor(0)= Treat XOStatus;
   if (XOTime = . or Time < XOTime) then XOStatus=0;
   else XOStatus= 1;
run;
```

The XOStatus variable takes the value 1 or 0 at time t (measured from the date of study treatment start), depending on whether or not the patient has received a t=crossover treatment at that time. Note that the value of XOStatus changes for subjects in each risk set (subjects still alive just before each distinct event time); therefore, the variable cannot be created in the DATA step.

3). Recurrent data – Multiple events

Many applications involve repeated events, where a subject may experience multiple events over a trial period or lifetime, such as myocardial infarction, infections, disability episodes, adverse events, hospitalizations. modeling time to first event only will not be adequate in this case. There is a growing interest in analysis of recurrent events data, also called repeated events data and recurrence data in clinical studies. The recurrent events data is ordered in a natural ordering of the multiple events within a subject.

Several methods have been presented on how to perform repeated events analysis.

- Intensity Andersen-Gill model
- Proportional rates and means model
- PWP (Prentice, Williams, Peterson) total time model
- PWP gap time model

AG model is a counting process approach where a subject contributes to the risk set for an event as long as the event occurs under observation and subject shares same baseline hazard function. It is a generalization of the Cox proportional hazards model and relates the intensity function of event recurrences to the covariates multiplicatively. It treats each subject as a multi-event counting process with essentially independent increments. To take into account the within subject correlation, the proportional means model is set up with sandwich variance estimate, which provides robust sandwich variance estimate for standard errors of coefficients, and does not require specification of the correlation matrix.

The following SAS statements provide an example of Intensity Model and Proportional Means Model:

```
proc phreg data=msurv covm covs(aggregate);
 model (Tstart,Tstop)*Status(0)=Trt;
 id usubjid;
run;
```

We first fit the intensity model to use model-based covariance estimate with **COVM** option. The **covs** option in proportional means model, requests a robust sandwich estimate for the covariance matrix which results in a robust standard error for the parameter estimates. A modified score test is also computed for testing the global null hypothesis. The aggregate keyword in the **covs** option requests a summing up of the score residuals for each distinct id pattern.

Conditional PWP models consider two time scales, a total time from the beginning of the study and a gap time from immediately preceding event. The PWP models are stratified Cox-type models that allow the shape of the hazard function to depend on the number of preceding events and possibly on other characteristics. Before use PROC PHREG for PWP models, the input data set has to be prepared to provide the correct risk sets. The input data set for analyzing the total time is the same as the AG model with an additional variable to represent the stratum that the subject is in. A subject with K events contributes K+1 observations to the input data set, one for each stratum that the subject moves to. To analyze gap time, the input data set should also include a GapTime variable that is equal to (TStop - TStart).

4). Frailty models – Incorporate longitudinal and survival data

In many clinical studies, longitudinal data and survival data frequently go hand in hand. Two typical examples are HIV and cancer studies. In HIV studies patients who have been infected are monitored until they develop AIDS or die, and they are regularly measured for the condition of the immune system using markers such as the CD4 lymphocyte count or the estimated viral load. Similarly in cancer studies the event outcome is death or metastasis and patients also provide longitudinal measurements of antibody levels or of other markers of carcinogenesis, such as the PSA levels for prostate cancer. These two outcomes are often separately analyzed using a mixed effect model for the longitudinal outcome and a survival analysis for the event outcome.

Longitudinal data and survival data are often associated in some ways. The time to event may be associated with the longitudinal trajectories. Separate analyses of longitudinal data and survival data may lead to inefficient or biased results. Joint models of longitudinal and survival data, on the other hand, incorporate all information simultaneously and provide valid and efficient inferences.

Individuals (with lower CD4 or higher PSA) are more frail than others, that is, the event in question is more likely to happen for them. One way of doing this is to assume that each individual has a frailty Z.
Then, conditional on the frailty, the hazard rate of an individual is assumed to take the form

$$\lambda(t|Z) = Z \cdot \lambda(t)$$

and the frailty model as

$\lambda i(t) = \lambda 0(t)\ eXi\beta+Zi\omega$, where Xi and Zi are the covariate matrices and ω is a vector of unknown random effects that describe excess risk or frailty.

Currently there is no direct SAS procedure to carry out the computations. SAS procedures PROC NLMIXED and PROC GLIMMIX can be used to jointly analyze a continuous and binary outcome.

5). Bayesian analysis – Bayesian modeling and inference

Bayesian analysis treats parameters as random variables and define probability as "degrees of belief" (that is, the probability of an event is the degree to which the event is true). Begin with a prior belief regarding the probability distribution of an unknown parameter when performing a Bayesian analysis. After learning information from observed data, the beliefs about the unknown parameter and obtain a posterior distribution will be updated. Bayesian methods offer simple alternatives to statistical inference—all inferences follow from the posterior distribution. However, most Bayesian analyses require sophisticated computations, including the use of simulation methods. MCMC(Markov chain Monte Carlo) simulation procedure fits Bayesian models with arbitrary priors and likelihood functions. The BAYES statement in PROC PHREG requests a Bayesian analysis of the regression model by using Gibbs sampling. Summary statistics, convergence diagnostics, and diagnostic plots are provided for each parameter. The following are the typical statements in Bayesian analysis.

```
ods graphics on;
proc phreg data=surv;
   model Time*Censor(0)= Treat ;
   bayes seed=1 outpost=Post;
run;
ods graphics off;
```

The BAYES statement invokes the Bayesian analysis. The SEED= option is specified to maintain reproducibility; the OUTPOST= option saves the posterior distribution samples in a SAS data set for post processing; no other options are specified in the BAYES statement. By default, a uniform prior distribution is assumed on the regression coefficient Treat. The uniform prior is a flat prior on the real line with a distribution that reflects ignorance of the location of the parameter, placing equal probability on all possible values the regression coefficient can take. Using the uniform prior in the following example, you would expect the Bayesian estimates to resemble the classical results of maximizing the likelihood. If you want to elicit an informative prior on the regression coefficients, you can use the COEFFPRIOR= option to specify it.

References

SAS Institute, Inc. SAS/STAT ® User’s Guide, SAS OnlineDoc® 9.22, Cary, NC: SAS Institute, Inc.

Contact Information

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