

Kaplan Meier And Cox Proportional Hazards Modeling: Hands On Survival Analysis

Tyler Smith, Department of Defense Center for Deployment Health Research,
Naval Health Research Center, San Diego, CA

Besa Smith, Department of Defense Center for Deployment Health Research,
Naval Health Research Center, San Diego, CA

ABSTRACT

Non-parametric survival analysis techniques are often used in clinical and epidemiologic research to model time at risk until event without parametric assumptions. This workshop will walk through the concepts of follow-up time, event time, the hazard function, the cumulative distribution function, incomplete data, censoring, time dependencies or temporal biases, plotting of survival curves, testing the proportional hazards assumption, and model diagnostics. Using SAS® system's PROC LIFETEST, Kaplan Meier curves along with the log rank and Wilcoxon tests will be investigated to establish statistical differences in survival times between two groups. From there we will use the SAS® system's PROC PHREG to run a Cox regression to model time until event while simultaneously adjusting for influential covariates and accounting for problems such as attrition, delayed entry, and temporal biases. The workshop will conclude with using the baseline option to calculate survival function estimates for graphing the cumulative probability of event over the follow-up period.

This workshop is aimed at intermediate level statisticians, epidemiologists, and data analysts.

INTRODUCTION

Survival analysis techniques employ methods designed to investigate the amount of study time an experimental unit contributes to a study period from entry until event. The term “survival” may be misleading because the techniques are applicable to any well-defined event although traditionally death was the event of interest and the study period consisted of following the subject until death. Events in survival analysis (also referred to as endpoints or outcomes) are defined by a transition from one discrete state to another at an instantaneous moment in time. Examples of events include months until onset of disease, days until remission after cancer therapy, years until stockmarket crash, hours until equipment failure, days until unemployment, or time until failing or passing an exam.

Although the origin of survival analysis goes back to mortality tables from centuries ago, recent advancements in survival analytic techniques using non-parametric and semi-parametric approaches have allowed researchers flexibility in their work not previously seen within the confines of parametric methods. These methods have become popular over parametric methods due to the relatively robust modeling approaches without distributional assumptions on the survival times.

Survival analysis has become a popular tool in observational and experimental studies involving follow-up of study participants over time. These studies often experience late arrival and early departure of subjects into and out of the observation period. Survival analysis techniques allow for a study to start without all experimental units enrolled and to end before all experimental units have experienced an event. This is extremely important because even in the most well developed studies, there will be subjects who choose to quit participating, move too far away to follow, die from some unrelated event, or will simply not have an event before the end of the observation period. With optional survival techniques, the researcher is no longer forced to withdraw the experimental unit and all associated data from the study. Instead, censoring techniques enable researchers to analyze incomplete data due to delayed entry or early withdrawal from the study. This is important in allowing each experimental unit to contribute all of the information possible to the model for the amount of time the researcher is able to observe the unit.

The ease with which survival analytic techniques have been applied in recent years owes much of its success to the availability of specialized software packages and high performance computing. Programmers and statisticians are now able to run complex and computationally intensive algorithms used in these types of analyses relatively quickly

and efficiently. However, these advancements have also made it possible for the novice to dabble in statistical environments where understanding of the fundamentals are essential. This Hands on Workshop will begin with univariate investigation of survival estimates using Kaplan-Meier curves and will conclude with adjusted hazard ratio estimates and survival curves using multivariable Cox Proportional Hazards regression.

TIME

The continuum that time reflects also implies that the probability of an event at an infinitely small single point in time is zero. Therefore it is necessary to define the distribution of events over that continuum instead of at an instant in time. In survival analysis, researchers rely on four functions to describe the distribution of event times: 1) probability density function (pdf), 2) cumulative distribution function (cdf), 3) hazard function, and 4) survival function.

These functions are quantitatively related to one another and possess a one-to-one relationship that makes interpretation and comparison easier. The pdf can be computed by taking the derivative of the cdf and likewise, the cdf can be computed by taking the integral of the pdf. The survival function is simply 1 minus the cdf, and the hazard function is calculated by dividing the pdf by the survival function. It is important to note that these relationships will allow us to compute the cdf from the survival function estimates produced by the SAS procedure PROC PHREG.

THE CUMULATIVE DISTRIBUTION FUNCTION

The cumulative distribution function is very useful in describing the continuous probability distribution of a random variable, such as time, in survival analysis. The cdf of a random variable T , denoted $F_T(t)$, is defined by $F_T(t) = P_T(T < t)$. This is interpreted as a function that will give the probability that the variable T will be less than or equal to any value t that we choose. Several properties of a distribution function $F(t)$ can be listed as a consequence of the knowledge of probabilities. $F(t)$ ranges from $0 < F(t) < 1$, is a nondecreasing function of t , and as t approaches ∞ , $F(t)$ approaches 1.

THE PROBABILITY DENSITY FUNCTION

The probability density function is also very useful in describing the continuous probability distribution of a random variable. The pdf of a random variable T , denoted $f_T(t)$, is defined by $f_T(t) = d F_T(t) / dt$. That is, the pdf is the derivative or slope of the cdf. Every continuous random variable has its own density function, the probability $P(a < T < b)$ is the area under the curve between times a and b .

THE SURVIVAL FUNCTION

Let $T > 0$ have a pdf $f(t)$ and cdf $F(t)$. Then the survival function takes on the following form:
 $S(t) = P\{T > t\} = 1 - F(t)$

The survival function gives the probability of surviving or being event-free beyond time t . Because $S(t)$ is a probability, it is positive and ranges from 0 to 1. It is defined as $S(0) = 1$ and as t approaches ∞ , $S(t)$ approaches 0. The survival curve describes the relationship between the probability of survival and time. Thus, $S(10)$ is the probability that an individual survives longer than 10 units of time, while $F(10)$ is the probability that an individual survives no more than 10 units of time.

THE HAZARD FUNCTION

The hazard function $h(t)$ is given by the following:

$$\begin{aligned} h(t) &= P\{t < T < (t + \Delta) \mid T > t\} \\ &= f(t) / (1 - F(t)) = f(t) / S(t) \end{aligned}$$

The hazard function describes the concept of the risk of an outcome (e.g., death, failure, hospitalization) in an interval after time t , conditional on the subject having survived to time t . It is the probability that an individual dies somewhere between t and $(t + \Delta)$, divided by the probability that the individual survived beyond time t .

The hazard function seems to be more intuitive to use in survival analysis than the pdf because it quantifies the instantaneous risk that an event will take place at time t given that the subject survived to time t . Sir David Cox recognized this appeal and in a sentinel paper published in 1972 described what is now known as the Cox Proportional Hazards model. In his paper titled, “Regression Models and Life Tables”, he outlines a robust regression method that did not require the choice of a probability distribution to represent survival times. We return for more description of this important paper later.

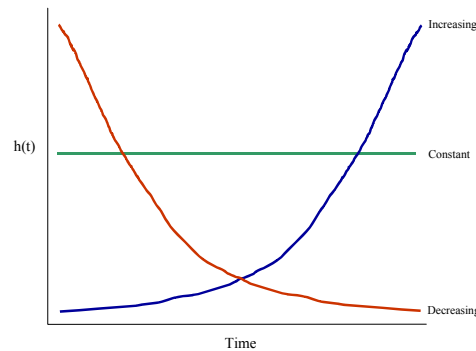


Figure 1. The plot of a constant hazard such as seen with accidents, an increasing hazard such as seen with the aging process of a mechanical engine, and a decreasing hazard such as seen with risk of dying after surgery.

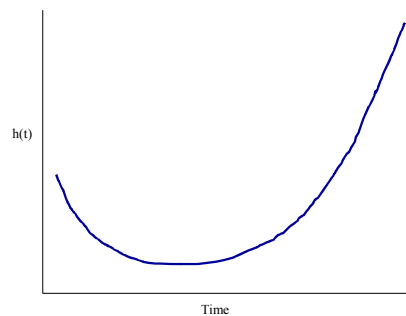


Figure 2. The plot of the hazard of death during a lifetime begins high at birth then goes down for many years before beginning to steadily increase through the aging process.

INCOMPLETE DATA

Although incomplete data plagues many analytic approaches, this unfortunate hurdle is common in survival analysis. A hallmark of survival analysis is the ability to manage many incomplete data forms. Survival analytic techniques rely on two points of observation time that must be carefully defined, as well as possible, prior to analyses. There is a beginning point of observation where time=0 and there is an ending point of observation where some reason or cause will terminate observation time. For example, in a complete observation cancer study, observation of survival time may begin on the day a subject is diagnosed with cancer and end when that subject dies as a result of the cancer. This subject is what is called an uncensored subject, resulting from the event occurring within the time period of observation. Complete observation time data like this example are desired but not realistic in most studies. There are always possibilities that the subject might recover completely and never have an event or the subject might die due to an entirely unrelated cause. In other words, the study cannot go on indefinitely waiting for an event from a participant, and unforeseen things happen to study participants that make them unavailable for observation. Specialized censoring techniques have been developed to ease the burden of these observation times.

LEFT AND RIGHT CENSORING

The most common form of censoring for incomplete data is right censoring where a subject's follow-up time terminates before the outcome of interest is observed. There are different forms of right censoring. The first, type I right censoring, occurs when the observation time reaches the end of a defined study period and the subject has not had an event. Type II right censoring occurs when the researcher ends the follow-up period based on a pre-specified number of events occurring. Right censoring also includes censoring a subject when they move out of observation during the follow-up period. Right censoring techniques allow subjects to contribute to the model until they are no longer able to contribute (end of the study, or withdrawal), but right censoring for loss to follow-up will be appropriate only if it is non-informative.

An observation is left censored if the event of interest has already occurred when observation of time begins. For example, in a study of myocardial infarction we begin following a group of people at age 50. However, some may have already had an event prior to the start of follow-up and unless you gain information as to the time of the events, the myocardial infarction may be left censored at age 50. In this paper we focus on the more typical right censoring.

The following figure presents a study design where the observation times start at differing points after the beginning of the study period. After $t=0$ is established, there is a fixed follow-up period. The X's represent events and the O's represent censored observations. Some subjects have events early in the study period and others have events at the end of the study period. Likewise some subjects enter the study period late and/or leave the study period early, but most do not have an event during the entire study and are simply right censored at the end. In this example there is no need for truncation techniques and we assume the censoring to be non-informative.

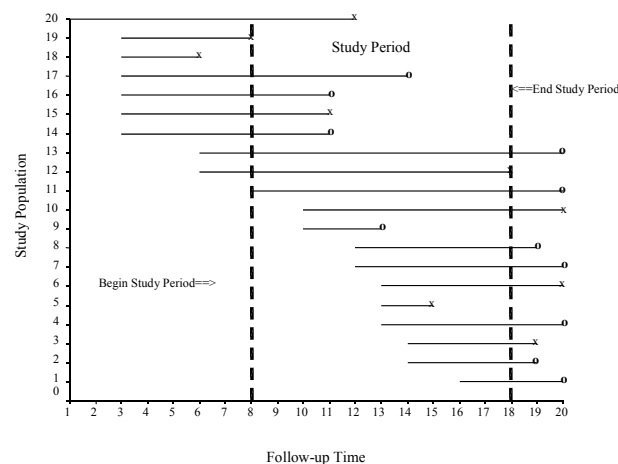


Figure 3. Follow-up time in a typical survival analysis.

KAPLAN-MEIER

The Kaplan-Meier (KM) estimator, or product limit estimator, is the estimator used by most software packages because of the simplistic step approach. The KM estimator incorporates information from all of the observations available, both censored and uncensored, by considering any point in time as a series of steps defined by the observed survival and censored times. When there is no censoring, the estimator is simply the sample proportion of observations with event times greater than t . The technique becomes a little more complicated but still manageable when censored times are included.

The KM estimator is a nonparametric estimator of the survivor function $S(t)$.

$$\hat{S}(t) = \prod_{t_s \leq t} \left(1 - \frac{d_j}{n_j} \right)$$

where d_j is the number of individuals who experience the event at time $t(j)$, and n_j is the number of individuals who have not yet experienced the event at that time and are therefore still at risk for experiencing it.

The KM estimator consists of the product of a number of conditional probabilities resulting in an estimated survival function $S(t)$ in the form of a step function. Using PROC LIFETEST we can compute and plot the survival curve of a single group or we can compare survival in subgroups. The procedure will also output tests for equality of the survival function estimates of the two or more strata being investigated.

```
proc lifetest data=analydat plots=(s) method=km graphics outsurv=surv ;  
  strata variable; * cutpoint if continuous;  
  time survtime*censor (value);  
  symbol1 line=1 color=blue;  
  symbol2 line=2 color=red;  
run;
```

PLOTS=(s) requests a plot of the survival step functions (curves).

METHOD=km requests the KM estimator (also the default) instead of the life tables method used when the researcher wishes event times to be grouped into intervals.

GRAPHICS requests high-resolution graphics rather than character based graphics.

OUTSURV=surv requests that a dataset be created with survival probabilities and confidence intervals.

STRATA statement indicates the group variable you wish to compare by. For categorical variables, just indicate variable name, for continuous indicate cutpoint—ex: (45) for age. The strata statement is left out if no comparisons are being calculated

survtime*censor (censor value) indicates the time of the event or censoring and whether or not the observation was censored.

symbol1 line=1 color=blue will color the lines in the graph on the output.

Plot Estimated Survival Curve of Cancer Patients After Surgery

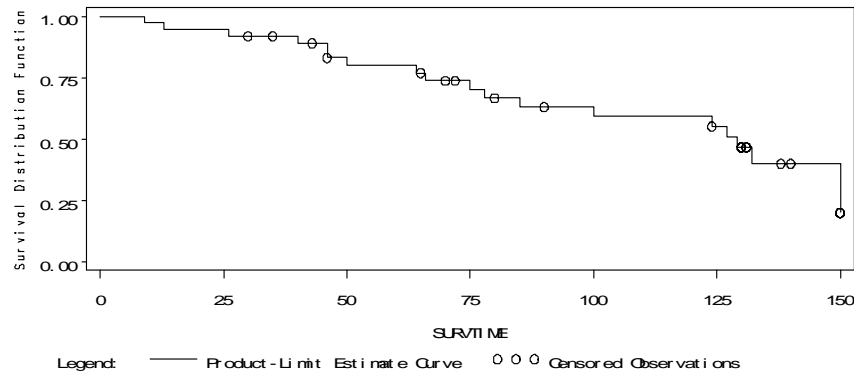


Figure 4. Kaplan-Meier survival step function (curve) of a single group.

PROC LIFETEST is often used to investigate the unadjusted survival times of a group without influence of other covariates in the model. It is also used as a non-parametric survival analysis approach when the proportional hazards assumption in Cox regression is violated. The procedure will output the mean, median, and quartile survival times of the group or subgroups of the population. The procedure will also report statistics for the comparison of two groups of survival times using the log rank test, Wilcoxon test, and the likelihood ratio test. The log-rank test, (which is the most commonly used and related to the Cox proportional hazards model) is more sensitive than Wilcoxon in detecting differences between groups occurring later in the follow-up. The Wilcoxon (better if no censoring) differs from the log-rank in that it takes into account the total number at risk at each time point. Neither test is good at detecting differences if the curves cross but they are adequate to compare two or more survival curves with the null hypothesis that all survival curves are the equal.

COX'S PROPORTIONAL HAZARDS REGRESSION

Sir David Cox's 1972 paper took a different approach to standard parametric survival analysis and extended the methods of the non-parametric Kaplan-Meier estimates to regression type arguments for life-table analyses. Cox advanced to prediction of survival time in individual subjects by only utilizing variables covarying with survival and ignoring the baseline hazard of individuals. Cox did this by making no assumptions about the baseline hazard of individuals and only assumed that the hazard functions of different individuals remained proportional and constant over time.

When there are several explanatory variables, and in particular when some of these are continuous, it is much more useful to use a regression method such as Cox rather than a KM approach.

Here the hazard function for individual i is modeled as:

$$h_i(t) = h_0(t)e^{\beta^T x_i}$$

where $h_0(t)$ is the baseline hazard function, β 's are regression coefficients, and x_i denote covariates.

The underlining or baseline hazard is the hazard when all covariates equal zero.

$$h(t, x) = h(t, 0)e^{\beta^T x}$$

$h(t,0)$ is the baseline hazard rate at time t for covariate vector 0. A subject's hazard at time t is proportional to the baseline hazard $h_0(t)$. The proportionality factor depends on the covariate vector for an individual. If all covariate values are homogenous, then it gets subsumed into the baseline hazard function.

The probability that an individual dies, leaves, etc., at time T_i , is given by:

$$\frac{e^{\beta x_j}}{\sum e^{\beta x_j}}$$

The conditioning eliminates the baseline hazard function.

Researchers favor Cox's proportional hazards modeling because of the robust semi-parametric method of calculating the probabilities of survival while simultaneously adjusting for other possibly influential variables. Other attractive features of Cox modeling include: the relative risk type measure of association, no parametric assumptions, the use of the partial likelihood function, and the creation of survival function estimates.

RELATIVE RISK TYPE MEASURE OF ASSOCIATION

The simple interpretation of the measures of association given by the Cox model as "relative risk" type ratios is very desirable in explaining the risk of event for certain categories of covariates or exposures of interest. For example, when a two-level (dichotomous) covariate with a value of 0=no and 1=yes is observed, the hazard ratio becomes e^{β} where β is the parameter estimate from the regression. If the value of the coefficient is $\beta = 1.099$, then $e^{1.099} = 3$. The measure is simply saying that the subjects labeled with a 1 (yes) are three times more likely to have an event than the subjects labeled with a 0 (no). In this way we have a measure of association that gives insight into the strength and direction of the relationship between our exposure and outcome.

NO PARAMETRIC ASSUMPTIONS

Another attractive feature of Cox regression is not having to choose the density function of a parametric distribution. This means that Cox's semi-parametric modeling allows for no assumptions to be made about the parametric distribution of the survival times, making the method considerably more robust. Instead, the researcher must only validate the assumption that the hazards are proportional over time. The proportional hazards assumption refers to the fact that the hazard functions are multiplicatively related. That is, their ratio is assumed constant over the survival time, thereby not allowing a temporal bias to become influential on the endpoint. In other words, the Cox proportional hazards model assumes that changes in the hazard of any subject over time will always be proportional to changes in the hazard of any other subject and to changes in the underlying hazard over time.

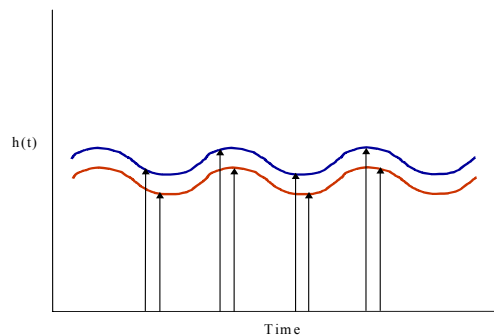


Figure 5. Graphical representation of proportional hazards over the follow-up period.

USE OF THE PARTIAL LIKELIHOOD FUNCTION

The Cox model has the flexibility to include time-dependent explanatory variables and handle censoring of survival times due to its use of the partial likelihood function. The likelihood function for the proportional hazards model can be thought of in two parts. The individual hazard and the exponentiated function of the independent variables represented by the linear sum of the $\beta_i x_i$'s. The baseline hazard multiplied by the function of independent variables produces the hazard for the i^{th} subject. The partial likelihood formula considers probabilities for those subjects who fail and does not explicitly consider probabilities for censored subjects. However, survival time information prior to censorship is used for those subjects who are censored. That is, the subject who is censored after the j^{th} failure time is part of the risk set used to compute the j^{th} likelihood even though this subject is censored later. We get estimates by finding values for the function of independent variables (betas) that maximize the partial likelihood. Some efficiency in estimate is lost but the model is robust and two of three standard properties of maximum likelihood estimates persist, being consistent and being asymptotically normal.

SURVIVAL FUNCTION ESTIMATES

With the SAS option BASELINE, a SAS dataset containing survival function estimates stratified by exposure levels can be output. These estimates correspond to the means of the explanatory variables for each stratum.

TIME DEPENDENCIES

In some situations the researcher may find that the dynamic nature of a variable causes changes in value over the observation time. In other instances the researcher may find that certain trends affect the probability of the event of interest over time. There are easy ways to test and account for these temporal biases within PROC PHREG but be careful if you have a large number of observations as the computation of the subsequent partial likelihood is very taxing and time consuming. Graphical display of survival curves is often an easier way to initially see if there are apparent proportional hazards assumption violations. If there is a steep increase or decrease in the survival curve, it may suggest more statistical investigation is needed. If you are investigating two or more survival curves and the curves cross, you need to investigate further. It is important to note here that when a time dependent variable is introduced into the model, the ratios of the hazards will not remain steady. This only affects the model structure. We will still be doing a Cox regression but instead the model used is called the extended Cox model.

The following is a quick note regarding extending the Cox model to incorporate time dependent covariates. The hazard at time t is the product of the baseline hazard function and the exponential expression is the linear sum of the $\beta_i x_i$'s. The baseline hazard is a function of time (t) but does not involve the independent variables, the function of independent variables involve the x 's but does not involve time (t). However, if we consider x 's that do involve t , we no longer have time independent x 's and must account for this in the interpretation of the model. The extended Cox model can be used with time dependent variables, but hazard ratios need to be interpreted as a function of time.

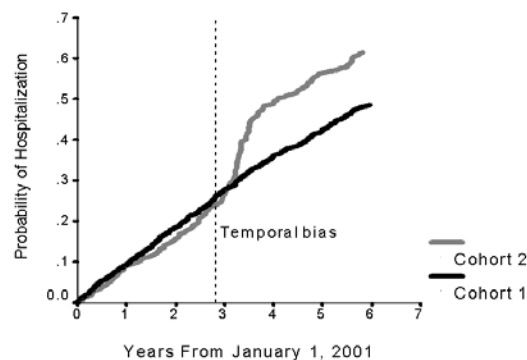


Figure 6. A cumulative distribution function that violates the proportional hazards assumption. Note the sharp increase in probability of hospitalization beginning right before the third year and lasting for approximately 1 year. After this one-year period the top curve then levels off and becomes parallel with the bottom curve once again.

ANALYTIC APPROACH

UNIVARIATE ANALYSES

Using PROC FREQ, PROC UNIVARIATE, and PROC LIFETEST an initial univariate analysis of the individual characteristics with event experience should be carried out to determine possible significant explanatory variables to be included in the model runs. An exploratory model analysis should be performed to explore the relations between the variables while simultaneously adjusting for all other variables that have influences on the event times. Collinearity among potential model variables should be investigated using PROC REG's diagnostic capabilities to ensure no model burdening correlations exist between variables. After investigation for confounding of variables not independently associated with the event times, variables with p-values of 0.05 or less are retained in the final model analysis. Additionally, the distributions of loss to follow-up times should be investigated for differing rates across the categories of exposure over the study period. The potentially harmful effect of differential loss to follow-up may be an indicator of informative censoring and should not be overlooked when conducting a survival analysis.

MULTIVARIABLE COX MODELING APPROACH

Dummy variables should be created using reference cell coding for the categorical variables. The measures of association output will compare your category of interest to the reference category of your choice. Starting with a saturated model, run PROC PHREG and use a manual backward stepwise model building approach. This will allow you to stay in control of the model and investigate possible confounding if variables are not independently associated with event times and slated for removal from the model. It is important to establish a magnitude difference rule such as 10-20% difference in measure of interest to establish confounding. Based on the rule you choose prior to beginning the analysis, retain variables that are not independently associated with the outcome but considered to be a confounder or remove from the model if considered not to be distorting the effect of interest. This will result in a final model with statistically significant independent risk factors of survival times or confounders.

```
proc phreg data=analydat;  
  model survtime*censor(0) (censor value) = treated x sex previousdxs / rl ties=efron;  
  x=treated*(log(survtime) - (log(mean survival)));  
  title1 'Cox regression of Treatment Status after Controlling for Sex and Previous Diagnosis';  
  title2 'Investigate Proportional Hazards Assumption';  
run;
```

DATA=analydat names the input data set for the survival analysis.

RL requests for each explanatory variable, the 95% (the default alpha level because the ALPHA= option is not invoked) confidence limits for the hazard ratios.

TIES=efron gives the researcher the approximations to the EXACT method without using the tremendous CPU it takes to run the EXACT method. Both the EFRON and the BRESLOW methods do reasonably well at approximating the EXACT when there are not a lot of ties. If there are a lot of ties, then the BRESLOW approximation of the EXACT will be very poor. If the time scale is not continuous and is therefore discrete, the option TIES=DISCRETE should be used.

x=treated*(log(survtime)-(log(mean survival))) tests the interaction of treatment with time to determine if the proportional hazards assumption is met. If x is not significant, you can conclude that the proportional hazards assumption is met and remove the variable from the model.

STRATIFICATION BY TREATMENT STATUS

These data were then stratified by treatment status in order to compute the survivor function estimates for the two treatment arms. Using the BASELINE function in PROC PHREG, we were able to output the survivor function estimates. The survival curves can then be displayed or we have the capability to compute the cumulative distribution function for the separate treatment arms over the study period.

```
proc sort data=survdat1;
  by treated;

proc phreg data=survdat1;
  by treated;
  model survtime*surv(0)= treated /rl ties=efron;
  baseline out=surv1 survival=s ;
  title1 'Cox Proportional Hazard Model';
  title2 'Survival Differences by Treatment';
  label
    treated = 'Yes Treated (no ref)';
run;
```

BY stratifies the analysis by the categories in the by variable, after data are sorted in that manner.

BASELINE without the COVARIATES= option produces the survival function estimates corresponding to the means of the explanatory variables for each stratum.

OUT=surv1 names the data set output by the BASELINE option.

SURVIVAL=s tells SAS to produce the survival function estimates in the output data set.

TEST statement allows testing of subgroups of regression coefficients. This statement is not shown above but can be done with “test age, occupation;” after the model statement. This test statement will test the null that age and occupation taken together are not related to probability of event after adjusting for the other variables in the model. This statement is also useful when testing the global significance of a categorical variable in which the model statement expresses only the dummy variables.

Graphing using PROC GPLOT

```
options ps=52;
goptions device=win;

symbol1 line=1 color=blue value=square i=join;
symbol2 line=2 color=red value=star i=join;

proc gplot data= surv1;
  plot survtime*s=treated;
  title1 font=swissb 'Cox Proportional Hazard Model' ;
  title2 font=swissb h=1.5 'Survival Differences by Treatment';
run;
```

Cox Proportional Hazard Model

Survival Differences by Treatment

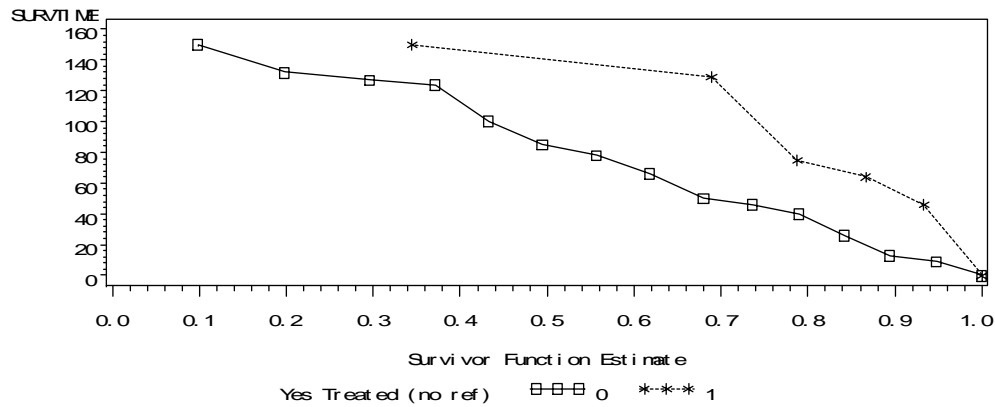


Figure 7. The stratified treatment arm survival curves for over the follow-up period.

Calculation of 1-survival function estimates in the output data set SURV1, obtained from running the BASELINE option, will produce the cumulative distribution function estimates that may be graphed as well (Figure 8).

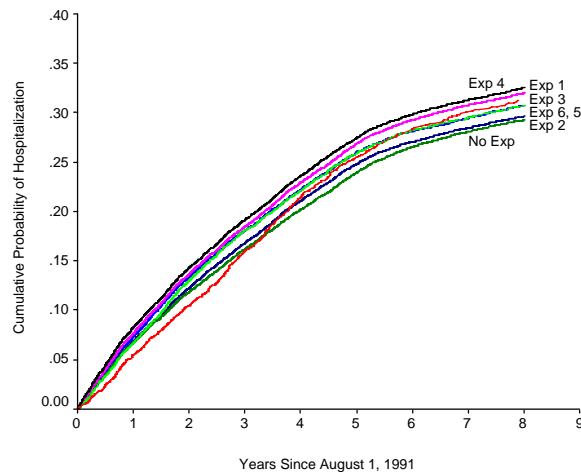


Figure 8: The stratified cumulative distribution functions of event by treatment arm. There was no violation of the assumption of proportional hazards but there did happen to be an observed significant difference in the probability of event between the 7 treatment arms.

LIKELIHOOD RATIO TEST

This test makes use of the log likelihood value given by the $-2\log L$ in the SAS output. If the researcher would like to see the importance of a variable or a group of variables in the model they should run a full and a reduced model. The full model has all of the variables included and the reduced model removes the variable or variables you would like to inspect. Taking the difference of the two values will yield a test statistic having a chi-square distribution under the null hypothesis with the number of degrees of freedom equal to the number of variables removed from the model.

COMPUTING THE GENERALIZED R^2

It may be helpful to compute the R^2 value for the Cox model. Although it is not an option of PROC PHREG, the R^2 value can be computed from the output of the regression.

$$R^2 = 1 - \exp(LR^2/n)$$

Where LR is the Likelihood-ratio chi-square statistic for testing the null hypothesis that all variables included in the model have coefficients of 0, and n is the number of observations. The researcher needs to take extreme caution when comparing the R^2 values of Cox regression models. Remember from linear regression analysis, R^2 can be artificially increased by simply adding explanatory variables to the regression model (ie; more variables does not equal a better model necessarily). Also, the above computation does not give the proportion of variance of the dependent variable explained by the independent variables as it would in linear regression, but does give a measure of how associated the independent variables are with the dependent variable.

RESIDUAL ANALYSIS

SAS has included three types of individual residuals for investigation. Cox-Snell residuals are helpful for assessing the fit of parametric models but are not as helpful with Cox models. Instead, the two that are most used are the martingale residuals and the deviance residuals. Deviance residuals are most often investigated because of their similarities with residuals from ordinary least squares (they are symmetrically distributed around 0 and have a standard deviation of approximately 1). Add the following after a model statement to output the martingale and deviance residuals:

```
baseline out=survs survival=s xbeta=xbet resmart=marting resdev=rdev;
```

The researcher can plot the residuals against covariates or plot the residuals against the linear predictor scores giving an idea of the fit or lack of fit of the model to individual observations.

```
proc gplot data=survs;  
  plot (marting rdev) * xbet / vref=0;  
  symbol1 value=circle;
```

MULTIPLE EVENT ANALYSIS

Often in long follow-up studies researchers will be asked if those treated have more events than those who were not treated. That is, have we wasted important event information by only modeling time until first event? There are many methods which can be used for this type of an analysis, however, the researcher first must decide if they are modeling ordered or unordered outcomes. Since we often investigate multiple events of the same type, we will focus on ordered outcomes. Three popular marginal regression models are the independent increment model (mutual independence of the observations within a subject), the WLW model (treating the ordered outcome data set with an unordered competing risks approach), and the conditional model (assuming an individual can't be at risk for outcome 2 if outcome 1 has not happened to the individual). After ordering the event dates in chronological order, follow-up time periods for multiple event time modeling are calculated from the start of follow up, and subsequent event dates, until first or subsequent event, loss to follow-up, or the end of follow-up, whichever occurs first. This area of survival analysis is currently of great interest among statisticians and epidemiologists and will no doubt have many advancements in the coming years.

SUMMARY

In this Hands on Workshop, we discussed the capabilities of SAS to allow the researcher to conduct a survival analysis using non and semi-parametric methods. The constellation of tools at the fingertips of SAS users include univariate procedures such as PROC FREQ, PROC TTEST, and PROC LIFETEST; regression diagnostic

procedures using PROC REG; the powerful multivariable procedure PROC PHREG; and the graphical procedure PROC GPLOT. With continued effort by SAS® to increase the capabilities of PROC PHREG this semi-parametric regression tool for handling incomplete data will only become more powerful. Further growth will hopefully include advanced graphics within PROC PHREG including the ability to output survival curves and the cumulative probability of event over the follow-up period. Using these analytic tools together is important to conduct a thorough analysis of time to event type data.

VERSION 9.0

Prior to SAS version 9.0 PROC PHREG did not have a class statement and the analyst was forced to create dummy variables. In addition, there was no automatic global test of the null hypothesis that a categorical variable was significant. With version 9.0 the PHREG procedure adds the class statement to the PHREG procedure with the idea that the enhancement will be incorporated into future PHREG releases. Also in version 9.0, the weight statement has been added and enables the analyst to specify case weights when using the BRESLOW or EFRON methods for handling ties. The test statement now includes the average option enabling the computation of a combined estimate of all the effects in the statement.

REFERENCES

Allison, Paul D., *Survival Analysis Using the SAS® system: A Practical Guide*, Cary, NC: SAS Institute Inc., 1995. 292 pp.

Cox DR. *Regression models and life tables* (with discussion). J R Stat Soc [Ser B] 1972;B34:187-220.

Hosmer JR. DW, Lemeshow S. *Applied Survival Analysis; Regression Modeling of Time to Event Data*. New York: John Wiley & Sons; 1999

Kaplan EL, Meier P. *Non-parametric estimation from incomplete observations*. J Am Stat Assoc 1958;53:457-481.
Kleinbaum DG, *Survival Analysis: A self-Learning Text*. New York: Springer-Verlag; 1996

SAS Institute Inc., *SAS/STAT® User's Guide, Version 6, Fourth Edition, Volume 1*, Cary, NC: SAS Institute Inc., 1989. 943 pp.

SAS Institute Inc., *SAS/STAT® User's Guide, Version 6, Fourth Edition, Volume 2*, Cary, NC: SAS Institute Inc., 1989. 846 pp.

SAS Institute Inc. *SAS/STAT® Software: Changes and Enhancements through Release 6.11*. Cary, NC: SAS Institute Inc., 1996. 1104 pp.

Smith TC, Heller JM, Hooper TI, Gackstetter GD, Gray GC. *Are Gulf War veterans experiencing illness due to exposure to smoke from Kuwaiti oil well fires? Examination of Department of Defense hospitalization data*. Amer J of Epidemiol; 2002; 155:908-17

Smith TC, Gray GC, Knoke JD. *Is systemic lupus erythematosus, amyotrophic lateral sclerosis, or fibromyalgia associated with Persian Gulf War service? An examination of Department of Defense hospitalization data*. Amer J of Epidemiol; 2000; Vol 151 No 11.

Gray GC, Smith TC, Knoke JD, Heller JM. *The postwar hospitalization experience among Gulf War veterans exposed to chemical munitions destruction at Khamisiyah, Iraq*. Amer J of Epidemiol; 1999; Vol 150 No 5.

Therneau TM, Grambsch PM. *Modeling survival data: extending the Cox model*. New York: Springer-Verlag; 2000.

ACKNOWLEDGMENTS

The authors would like to thank CDR Margaret AK Ryan, Director of the Department of Defense Center for Deployment Health Research at the Naval Health Research Center, San Diego.

Approved for public release: distribution unlimited.

This research was supported by the Department of Defense, Health Affairs, under work unit no. 60002.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

ABOUT THE AUTHORS AND CONTACT INFORMATION

Tyler has used SAS for 12 years as a senior statistician with the DoD Center for Deployment Health Research. Responsibilities include management, analysis, and interpretation of large demographic, health care, and longitudinal data on US military personnel. This work has culminated in more than 30 peer reviewed journal manuscripts in major scientific journals. Invitations to speak include the International Biometrics Society Meetings, and many invitations to the local San Diego group, WUSS, and SUGI conferences.

Tyler C. Smith, MS
Statistician, Henry Jackson Foundation
Department of Defense Center for Deployment Health Research, at the Naval Health Research Center, San Diego
smith@nhrc.navy.mil

Besa has been using SAS for 8 years including her work currently as a senior biostatistician with the DoD Center for Deployment Health Research at the Naval Health Research Center, San Diego. Her responsibilities include management of large military and demographic data sets, mathematical modeling and statistical analysis for longitudinal and health-based studies. She has presented at previous WUSS and local San Diego SAS user group meetings.

Besa Smith, MPH
Biostatistician, Henry Jackson Foundation
Department of Defense Center for Deployment Health Research, at the Naval Health Research Center, San Diego
besa@nhrc.navy.mil