Abstract
The analysis of survival-time data often involves an examination of relationships that are complex and difficult to display. This is particularly true when there is more than one covariate (independent variable) or when one of the covariates is continuous. Some examples of displays that are frequently needed but for which programming is tedious are:
1. Plots of Kaplan-Meier [1] 2-and 4-year survival probabilities vs. sex and deciles of age.
2. Plots of one or more survival curves (with followup time on the x-axis) stratified by quartiles of blood pressure and adjusted for 5 other covariates.
3. Plots of Cox [2] survival estimates vs. a continuous range of blood pressure, drawn separately by sex group.
OS SAS currently has the tools for generating such plots (e.g., PROCs PHGLM, RANK, GPLOT), but they may be cumbersome to use. This paper discusses a new SAS macro titled SRVTREND that is useful for displaying survival trends versus one, two or three covariates (optionally adjusted for other covariates), each covariate with its own assumption level.

Introduction

Let $T_i$ denote the survival time (failure time) and $X_i = [X_{i1}, X_{i2}, ..., X_{ip}]$ denote a vector of $p$ covariate values (predictor variables) for the $i$th individual. Of primary interest in survival analysis is the estimation of the survival function or the probability of surviving past time $t$ given covariate values $X_i$:

$$S(t|X_i) = Pr(T_i > t | X_i)$$

The form and strength of the relationship between $X_i$ and $S(t|X_i)$ is also of major interest.

The three most commonly used methods for estimating $S(t|X_i)$ are the following:

1. **Empirical:** Stratify and cross-classify the $X_i$ into groups and compute Kaplan-Meier [1] estimates of $S(t|X_i)$. No assumptions are made except that the groupings on $X_i$ are meaningful.

2. **Cox Model:** Making assumptions of proportional hazards over time and linear regression in the covariates, the model

$$S(t|X_i) = S_0(t) \exp(\beta_{X_{i1}} + ... + \beta_{X_{ip}})$$

is fitted by estimating regression coefficients using the maximum partial likelihood method [2]. No grouping of data is required, and no functional form is assumed for $S_0(t)$, the "underlying" survival function.

3. **Mixed Empirical/Cox Model:** By stratifying on some of the covariates, Cox-Kalbfleisch-Prentice survival estimates [3] can be obtained wherein PH is assumed with respect to some covariates, but nothing is assumed for the relationship between other covariates and survival. For example, suppose that one wants to examine the relationship between $X_i$ and $S(t|X_i)$ without making any assumptions with respect to $X_i$, while assuming PH with linearity for $X_{i1}$. Let the possible values of $X_{i1}$ be denoted by $v_1, v_2, ..., v_c$. The stratified Cox model is then

$$S(t|X_{i1} = v_j, X_{i2}, ..., X_{ip}) = S_jo(t) \exp(B_{v_j} + X_{i1}B_{v_1} + ... + X_{ip}B_{v_p})$$

where $S_jo(t)$ is the standard or "underlying" survival function for the $j$th stratum of $X_{i1}$. Survival estimates obtained using this model can be thought of as Kaplan-Meier/Cox estimates or Cox-adjusted Kaplan-Meier estimates. They are useful for examining the effect of $X_{i1}$ on survival while making assumptions only for the other covariates being "adjusted for". This method does have drawbacks, namely deciding how to categorize a continuous covariate $X_{i1}$ and choosing $x$-coordinate positions for graphing the regression function.

Grouping Continuous Covariates

When using stratification to obtain empirical survival estimates, the analyst must choose a method of grouping covariate values. When the covariate has few levels, it may be treated as a discrete variable and each possible value can be treated as one stratum. Otherwise, the covariate must be forced to become discrete by grouping it into intervals. Often, intervals are defined by rounding the covariate, but quantile grouping is usually preferred. For example, age may be grouped into quartiles to ensure equal numbers of observations in each of four groups. Interval widths will then vary from group to group so that adequate subsample sizes will be obtained, thus avoiding an attempt to estimate survival in a group containing too few observations.
Except when covariate values are discrete, the analyst must decide which x-coordinates to use for representing intervals, as this choice will affect the shape of the estimated regression function when the covariate is plotted on the x-axis. Many analysts use the interval midpoint for this purpose, but this may not be appropriate when the distribution of covariate values is highly skewed in an interval (as is typically the case in the lowest and highest intervals). Instead, the mean covariate value in the interval better represents that interval. We wrote a general statement-form macro called QUANTREP to replace covariate values with the mean of all values in a given quantile group. This process effectively discretizes a continuous variable and labels it with the mean value in each group so that the shape of the regression function may be estimated, without making the PH assumption, through the use of stratified Kaplan-Meier estimates.

For example, to group age into quintiles and the value of the variable age with the mean age in each quintile, we use the statement

```
QUANTREP age age groups=5;
```

Now age can be used as if it were a discrete variable having 5 levels. To form quintile groups independently for each sex group, the phrase CLASS=sex would be added to the statement above.

Cubic Spline Functions: Relaxing the Linearity Assumption

Frequently the analyst is willing to make the PH assumption (or has verified that the effect of a covariate on the hazard ratio is constant over time using the PLOTHR macro [4], for example), but at a given stage of the analysis is unwilling to assume that the linear regression assumption holds. By using a spline function to represent the effect of a covariate on log hazard or log-log survival, the analyst can model the effect of a continuous variable using all the power of the Cox model without assuming linearity of the regression function and without arbitrarily grouping the variable. A plot of the fitted spline function yields an estimate of the transformation needed so that the covariate would satisfy the linear regression assumption.

A cubic spline function (a piecewise cubic polynomial with smooth transitions at join points or "knots") restricted to be linear in the tails (beyond the outermost knots) has been proposed by Stone and Ko [5] and has been applied to logistic regression modeling in SAS by Devlin and Weeks [6]. This function will fit a variety of shapes, and it is easily fitted in the context of many regression models by setting up an appropriate design matrix, i.e., by deriving constructed variables involved in the truncated power basis formulation [6,7]. We wrote a macro procedure DASPLINE to automatically construct symbolic formulas for these derived variables, based on user-specified or automatically chosen knots using selected quantiles of the covariate's distribution.

DASPLINE is useful in any regression model that is linear in the regression coefficients. An example demonstrates how it is invoked:

```
DASPLINE "age bp1;"
DATA; SET:
 & age; *Derives age1-age3;
 & bp; *Derives bpl-bp3;
PROC REG;
MODEL y=age age1-age3
 bp bpl-bp3 ;
```

Here the factors age, age1, age2, and age3 are the original and three constructed variables needed to estimate one parameter for the spline function having 5 knots at default percentiles. Note that a test of significance for age1-age3 is a 3 degree of freedom test of linearity of the age effect. The two outermost percentiles to use for knots may be specified using parameters OUTER and SECOND. When SRVTREND calls DASPLINE, it uses the default percentiles 5, 25, 50, 75, 95.

The SRVTREND Macro Procedure

SRVTREND is a SAS statement-form macro language procedure that contains 500 SAS and SAS macro language statements. The procedure creates many types of survival trend plots by varying up to three covariates along with followup time. Survival probability or -log-log survival probability appears on the y-axis, and followup time or a covariate appears on the x-axis. Of the 1-3 factors used for the displays, one factor may describe what makes up different plots (e.g., a separate page for each sex group), one may describe different curves on one plot (e.g., a separate curve for each quartile of age), and one may describe the x-axis (if followup time is not on the x-axis). If a covariate is shown on the x-axis, survival at one or more specified time points is shown on the y-axis.

A powerful feature of SRVTREND is its ability to handle different levels of assumption for different covariates. For example, one factor for sex interaction, no assumption is made regarding the effect of sex on survival, while proportional hazards with linear regression is assumed for age. Also, the analyst may obtain what are essentially adjusted Kaplan-Meier survival curves for males and females, adjusted for age using a Cox model stratified by sex. Other than assuming no age x sex interaction, no assumption is made regarding the effect of sex on survival, while proportional hazards with linear regression is assumed for age. Another covariate could be added to the picture - one that is modeled with a spline representation so as not to assume linearity. A covariate such as one measuring severity of symptoms (mild, moderate, severe) could be modeled with two dummy variables so that the three survival curves have the same relative shapes but no assumption is made about the category spacings. By varying the level of assumption made for a predictor variable and comparing survival estimates, the analyst can learn much about the underlying data structure.

The assumption levels for each factor are coded as follows:

- 0: No assumption
- 1: Cox proportional hazards
- 2: Cox proportional hazards with linear regression
- 3: Cox proportional hazards with restricted cubic regression
- 4: Cox proportional hazards with truncated power basis
- 5: Cox proportional hazards with cubic spline

This allows for a wide variety of analyses and comparisons, providing a flexible tool for survival trend analysis.
<table>
<thead>
<tr>
<th>Level</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>No assumption. Round the variable if necessary or group it into quantiles and subgroup on its levels. This is the default assumption level.</td>
</tr>
<tr>
<td>1</td>
<td>Proportional hazards with no assumption regarding shape of regression function (uses dummy variables indicating different values). This level is used for modeling a discrete ordinal or nominal covariate such as one defining 4 treatments.</td>
</tr>
<tr>
<td>2</td>
<td>Proportional hazards without assuming linearity (uses restricted cubic spline regression function). This level is used mainly for continuous covariates.</td>
</tr>
<tr>
<td>3</td>
<td>Proportional hazards with linearity.</td>
</tr>
</tbody>
</table>

The type of model actually fitted by SRVTREND depends on the assumption levels of the covariates and on the presence of adjustment variables, as shown in the following table [the models are numbered according to the methods listed in the Introduction section and A.L. = assumption level].

<table>
<thead>
<tr>
<th>A.L.</th>
<th>A.L. &gt; 0</th>
<th>Adj.</th>
<th>Model Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>Yes</td>
<td>No or Yes</td>
<td>2</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>1</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>3</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>3</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>3</td>
</tr>
</tbody>
</table>

The SAS PHGLM procedure [8] is called from SRVTREND and used to estimate any regression parameters and obtain survival estimates. The BLOCK option is used if any assumption level 0 variables are present. If no variables with assumption level >0 are present and no adjustment variables are specified, there will be no covariates in the model, only stratification (blocking) factors. In addition to PHGLM, SRVTREND uses SAS procedures FREQ, UNIVARIATE, CONTENTS, SORT, FORMAT, PRINT, PLOT, and GPLDT (with the Annotate facility). SRVTREND calls the QUANTREP macro procedure if any quantile grouping is used, and the DASPLINE macro if any spline functions are used.

The user may specify a variable used to create separate plots (the P variable), a variable used to create separate curves on each plot (the CLASS variable), and a variable to display on the x-axis instead of followup time (the X variable). In addition, an optional list of variables to adjust for but not display may be given. SRVTREND is similar in spirit but is more general than the EMPTREND macro procedure used for examining assumptions of logistic and linear models [9].

SRVTREND is invoked by the following statement:

```
SRVTREND TIME = time variable
EVENT = event/censoring indicator
```

The options that may appear in the SRVTREND statement follow, with default values in parentheses.

```
DATA = input dataset (last one created)
P = variable describing different plots (none)
PASSUME = assumption level for P (0)
PRANGE = settings for P of the form PRANGE = "0 to 20" (default increment is 1)
PRANGE = "0 to 15 by 5"
PRANGE = "4, 5, 8" (must be of this form for assumption level 1)
PRANGE is mandatory for assumption level >1 and is not used for assumption level 0. It is recommended for assumption level 1 as this will save a PROC FREQ to determine the possible values of the covariate.

PGROUPS = Number of quantile groups to create (if PASSUME=0)
PROUND = round P to nearest r (if PASSUME=0)
CLASS = variable describing different curves on one plot (none)
CASSUME = CRANGE= CGROUPS= CROUND= as with P
X = variable on X-axis (default is followup time, X is used only if T is specified)
XASSUME = XCRANGE= XGROUPS= XROUND= as with P
NMIN = minimum stratum size to use in computing survival estimates if any assumption level 0 variables are used (10)
T = "settings for TIME" for computing survival probabilities if X is specified. Omit T and X to plot the entire survival curve over the followup period. T may have one of the following forms:
T="2" plot 2-UNIT survival only
T="2, 5" plot 2 and 5-UNIT survival
T="1 to 7 by 2" plot 1, 3, 5, 7-UNIT survival
T="2 to 4" plot 2, 3, 4-UNIT survival
UNIT = unit of measure for TIME (year)
ADJ = optional list of numeric variables to adjust for, enclosed in quotes if there are more than one. The variables not listed in ADJTO will be adjusted to the grand mean. Proportional hazards with linearity is assumed for the adjustment variables
```
ADJTO = optional settings for ADJ variables 
if you do not wish to use the grand 
mean, e.g. ADJTO="age=50;sex=1;" May 
contain formulas - 
ADJTO="age=50;agex=age*sex;" Note: 
ADJTO variables must appear in the 
ADJ, P, X, or CLASS parameters, and 
ADJTO variables assigned constant 
values must precede their use in 
formulas as illustrated above. To 
adjust to stratum-specific medians, 
first compute the medians and then 
use e.g.: CLASS=sex ADJ=weight 
ADJTO="weight=125*(sex='F') + 
180*(sex='M');" 

LOGLOG to obtain a plot for -log-log 
survival also 
LOGLOG=2 to plot -log-log axis in reverse 
order 
HAXIS = "optional x-axis specs for PLOT 
statement", e.g. HAXIS="0 to 100 BY 
10"
VAXIS = "optional y-axis specs for PLOT 
statement", "0 to 1 BY 1"
LAXIS = "optional y-axis specs for log-log 
survival!"
YLABEL = Y-axis label for survival plots 
("Survival Probability")

Legend label if CLASS used and 
PLOT=I, omit to use variable name, 
NONE to suppress legend label, "any 
label" to specify your own label. 
Applies only to log-log plots since 
class labels are placed besides 
regular survival curves.

PLOT = plotting option (I) 
1 = line printer 
2 = graphics device using SAS/GRAPH 
3 = both line printer and graphics 
device 
PNFOL = u to print on the graph the number 
of persons followed at least 0, u, 
2u, 3u, 4u, ... time units if time is 
on the x-axis (X not specified). 
Specify PNFOL=O to suppress 
printing (I)

Notes: The GROUPS and ROUND parameters are 
mutually exclusive. If the assumption level for 
any variable is 0 and the variable is discrete, 
ROUND and GROUPS may be omitted and the variable 
is also allowed to be a character variable. 
Assumption level I variables may also be 
character. This macro invokes the QUANTREP 
macro if a GROUPS parameter is used. When 
quantile grouping is used, each quantile is 
identified by the mean value of the variable in 
the quantile group. When time is on the x-axis, 
specifying HAXIS = "low TO high BY inc" will 
ensure that step curves are carried all the way 
to "high" if followup permits. Otherwise, the 
I=STEPLO on the SYMBOL statement will not carry 
the last step to the end of the graph. When a 
CLASS variable is present, CLASS labels are 
limited to 12 characters in length. P, CLASS, 
and more than one T point may not be specified 
simultaneously. If T is given, X must be given, 
and vice-versa. If CLASS and more than one T 
point are given but P is omitted, a separate 
graph will be produced for each time point. If 
more than one time is given but CLASS is not, a 
separate curve will appear on each graph for 
each value of time.

The various options are best understood 
through examples. Examples 1 and 2 demonstrate 
analyses of secular trends using actual data 
from the Duke Cardiovascular Disease Databank. 
Examples 3-11 use simulated exponential survival 
data with various hazard functions h(x) that are 
independent of time t, implying that the true 
population -log-log S(tlx) is -log(t) -log h(x) 
and proportional hazards holds. In these 
examples one can compare plotted estimates with 
true, known functions. If the true regression 
function is linear, log h(x) is a linear 
function of x.

Example 1: Subset on Calendar Year (yr, an 
integer, not to be confused with followup time) 
and Treatment (surg). Plot adjusted 
Kaplan-Meier 2-year survival versus yr adjusted 
for ef and sy, where ef is to be adjusted to 50 and sy is 
calendar year by treatment interaction.

Example 2: Stratify on treatment (surg; no PH 
assumption). Plot Cox predicted survival curves 
for Calendar years 70 and 77 after fitting a 
linear secular trend. Adjust for ef and sy, 
where ef is to be adjusted to 50 and sy is 
calendar year by treatment interaction.
Example 3: Simulated exponential survival data.
Generate exponentially distributed survival time data with hazard \( h(x) = 0.02 \exp(3|x - 0.5|) \), where \( x \) is \( U(0,1) \). Censoring distribution is \( U(0,15) \).

SAS Program for examples 3-5.

```sas
DATA test;
    Do i=1 TO 3000;
    cens=15*RANUNI(12); x=RANUNI(9);
    h=0.02*exp(3*abs(x-.5));
    t=-log(RANUNI(5))/h;
    If t<=cens THEN e=1;
    ELSE Do; e=0;
    t=cens; END;
    OUTPUT;
END;
TITLE "h(x) = 0.02*exp(3|x-.5|)" ;
```

Plot 5 and 8 year Cox model predicted survival curves versus \( x \). Assume PH without assuming linearity.

SRVTREND TIME=t EVENT=e X=x XASSUME=2
XRANGE="0 TO 1 BY .01" T="5,8" PLOT=2;

Example 4: Example 3 with no assumptions (Kaplan-Meier 5 and 8 year survival).

SRVTREND TIME=t EVENT=e X=x XGROUPS=10
T="5,8" LOGLOG PLOT=2;

```
```
Example 5: Example 3 inappropriately assuming PH with linearity.

SRVTREND TIME=t EVENT=e X=x XASSUME=3 XRANGE="0 TO 1 BY .01" T="5,8" PLOT=2;

\[
h(x) = 0.02 \cdot \exp(3x - 5)\]

Example 6: Simulated exponential survival data. Generate exponentially distributed survival time data with \( h(x) = 0.02 \cdot \exp(0.04 \cdot (\text{age} - 50) + 0.8 \cdot \text{sex}) \), where \( \text{age} \) is \( \mathcal{N}(50,144) \). Censoring distribution is \( U(0,15) \).

SAS Program for examples 6-8.

```sas
PROC FORMAT;
VALUE MF 0="MALE" 1="FEMALE"
DATA test;
DO i=1 TO 2000;
AGE=50+12*RANNOR(13);
SEX=I*(RANUNI(7) <= 0.4)
cens=15*RANUNI(6);
h=0.02*exp(0.04*(AGE-50)+0.8*SEX);
t=-log(RANUNI(4))/h;
IF t<=cens THEN e=1 ELSE DO; e=0; t=cens; END;
OUTPUT;
END;
FORMAT SEX MF. ;
TITLE "h(x)=0.02*exp(0.04*(age-50)+0.8*sex)";
```

Stratify on age and sex. Plot 3 year Kaplan-Meier survival estimates versus sex group and deciles of age.

SRVTREND TIME=t EVENT=e X=age XGROUPS=10 T=3 CLASS=sex LOGLOG PLOT=2;

Example 7: Example 6 assuming PH with linearity.

SRVTREND TIME=t EVENT=e X=age XASSUME=3 XRANGE="30 TO 70 BY 2" T=3 CLASS=sex LOGLOG PLOT=2;

\[
h(x) = 0.02 \cdot \exp(0.04 \cdot (\text{age} - 50) + 0.8 \cdot \text{sex})\]
Note that Examples 7 and 8 demonstrate the gain in precision/stability of estimates over Kaplan-Meier estimates resulting from making (true) model assumptions; the model \(-\log(\log(t))-\log(0.02 - [0.04(age-50) + 0.8sex])\).

Example 9: Simulated exponential survival data. Generate exponentially distributed survival time data with \(h(x) = 0.02 \exp(0.7\log(x))\), where \(\log x\) is \(N(2,1)\). Censoring distribution is \(U(5,15)\).

SAS Program for examples 9-11.

```sas
DATA test;
DO i=1 TO 3000;
cens=5+10*RANUNI(5); X=exp(2+RANNOR(6));
h=.02*exp(0.7*log(X));
t=-log(RANUNI(6))/h;
IF t<=cens THEN e=1; ELSE e=0; t=cens;
END;
OUTPUT;
END;
TITLE "h(x)=0.02*exp(0.7*log(x))";
```

Plot 8 year Kaplan-Meier survival estimates versus deciles of \(x\).

```sas
SRVTREND TIME=t EVENT=e X=x XGROUPS=10
HAXIS="0 TO 125 BY 25" T=8 LOGLOG PLOT=2;
```

```
h(x)=0.02*exp(0.7*log(x))
```
Example 10: Example 9 incorrectly assuming PH with linearity.

SRVTREND TIME=t EVENT=e X=x XASSUME=3
XRANGE="0 TO 125 BY 1" HAXIS="O TO 125 BY 25"
T=8 PLOT=2 ;

Note the agreement between the function plotted in the last figure with the logarithmic regression function \(-\log h(x) = -\log 0.02 - 0.7 \log(x)\).

Example 11: Example 9 assuming PH without assuming linearity.

SRVTREND TIME=t EVENT=e X=x XASSUME=2
XRANGE="0 TO 125 BY 1" HAXIS="O TO 125 BY 25"
T=8 LOGLOG PLOT=2 ;

Note the agreement between the function plotted in the last figure with the logarithmic regression function \(-\log h(x) = -\log 0.02 - 0.7 \log(x)\).
Summary

The SRVTREND procedure serves three major functions. First, it allows the analyst to perform a rather complex survival analysis with only a minimum of program specifications. Second, it makes presentation-ready graphics displaying various kinds of survival probability estimates, including such features as plotting the number of individuals at risk during each follow-up interval using the SAS/GRAPH Annotate facility. Third, the procedure ties together many different components, such as a) quantile grouping for factors the analyst does not wish to model (i.e., for deriving stratified Kaplan-Meier estimates), b) automatic generation of design matrices for fitting restricted cubic splines for continuous covariates for which the analyst is unwilling to assume linear regression, c) maximum likelihood estimation of regression parameters and survival probabilities (at specified time points or over the entire follow-up period), and d) graphing the predicted survival probabilities versus 1-3 factors. In addition to displaying data and model results, SRVTREND is useful for checking Cox model assumptions and for graphing needed covariate transformations.

Availability of Software

Copies of SAS macros described in this paper may be obtained free of charge if you are served by the BITNET electronic mail facility (contact DFRANK@TUCC). Alternatively, write to Duke University Medical Center, Box 3363, Durham, NC 27710 for information on obtaining the software.

Acknowledgements

This work was supported by Contracts N43-CM67949 and N43-CM67945 from the National Cancer Institute, and Research Grants HS-04873 and HL-17670 from the National Center for Health Services Research and the National Heart, Lung, and Blood Institute, respectively.

References


SAS and SAS/GRAPH are registered trademarks of SAS Institute, Inc., Cary, N.C., U.S.A.