

Surviving Survival Analysis – An Applied Introduction

Christianna S. Williams, Abt Associates Inc, Durham, NC

ABSTRACT

By incorporating time-to-event information, survival analysis can be more powerful than simply examining whether or not an endpoint of interest occurs, and it has the added benefit of accounting for censoring, thus allowing inclusion of individuals who leave the study early. This tutorial-style presentation will go through the basics of survival analysis, starting with defining key variables, examining and comparing survival curves using PROC LIFETEST and leading into a brief introduction to estimating Cox regression models using PROC PHREG. The evaluation of the proportional hazards assumption and coding of time-dependent covariates will also be explained. The emphasis will be on application, not theory, but pitfalls the analyst must watch out for will be covered. Examples will be taken from real-world data from health research, and some features newly available in SAS® 9.2 will be highlighted.

INTRODUCTION

Broadly speaking, survival analysis is a set of statistical methods for examining not only event *occurrence* but also the *timing* of events. These methods were developed for studying death – hence the name survival analysis – and have been used extensively for that purpose; however, they have been successfully applied to many different kinds of events, across a range of disciplines. Examples include manufacturing or engineering: how long it takes widgets to fail; meteorology: when will the next hurricane be hit the North Carolina coast; social: what determines how long a marriage will last; financial: the timing of stock market drops – the list goes on. Sometimes other names are used to refer to this class of methods – such as event history analysis, or failure time analysis or transition analysis, but many of the basic techniques are the same as is the underlying idea – understanding the pattern of events in time and what factors are associated with when those events occur.

Of course, books have been written on this topic – a couple are even listed at the end of this paper – and I have neither the time, nor the space – nor the competence – to describe all aspects of survival analysis – or even all the SAS survival analysis methods. Further, this paper is not intended to explain the statistical underpinnings of survival analysis. Rather, it is my intent to go through the analysis of one set of data in some detail, covering many of the basic concepts and SAS methods that the programmer/analyst needs to know. I want to give you an intuitive sense of how some basic survival analysis techniques work, and how to write the SAS code to implement them. Also, the last few releases of SAS, including 9.2, have some great new features for the survival analysis procedures – I will give you a taste of those too. The specific topics to be covered include:

- Creating the survival time and censoring variables – the good old DATA step;
- A fairly detailed treatment of Kaplan-Meier survival curves; overall and stratified, as implemented in PROC LIFETEST; and
- A brief introduction to Cox Proportional hazard models (PROC PHREG), including a few comments on proportionality and the coding of time-dependent covariates.

I'll also be upfront about some of the topics I am not going to cover. I'm not going to give more than a passing mention to the following: parametric survival analysis (e.g. PROC LIFEREG), recurrent events, left or interval censoring, Bayesian methods. Many of the more advanced features in PHREG will also not be addressed. I am also not going to talk about ODS graphics with respect to LIFETEST...though I encourage you to explore!

GETTING STARTED

A schematic depiction of simple survival data for six subjects is shown in Figure 1. In this figure, all subjects start their survival time at the same point – the study baseline. Further, we assume that each person can have the event only once. Three of the six patients (lines ending in solid circles -- #1,3, and 6) have an “event”, and we can ascertain how long each of them was in the study prior to their event – their “survival time”. As noted above the event may be death, but it can also be any other endpoint of interest, where we can measure the date of onset. In the study from which the examples in this paper will be drawn, the outcome event of interest is nursing home admission.

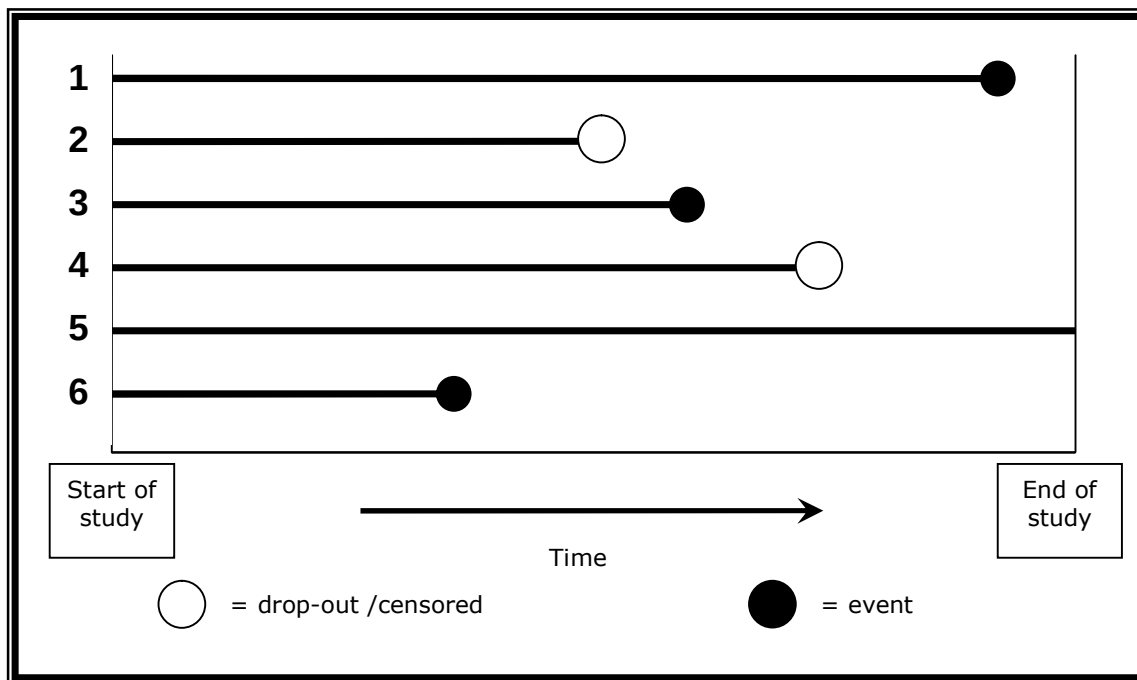


Figure 1. Hypothetical survival data for six patients. See text for further description.

In the Figure, there also 3 subjects (#2, 4 and 5) who do not have an event – at least not while they are in the study. Subject #5 is the only one who completed the entire study without having an event. In contrast, two of the cases (open circles, #2 and 4) are lost to the study before having an event and before the study follow-up ends; they are said to be **censored**. Actually, #5 is censored also – in this context, censoring simply means that at the end of a given individual's follow-up (whether that was early or at the end of the study), he/she had not had the event of interest. Different things can cause censoring, depending on the study design. It may be that these study participants decided they did not want to continue in the study, and so all we know is that – at the time they left the study, they had not yet had the event of interest. If our event of interest is not death, then it may be that censoring is caused by death – again, we know that at the time we stopped following that person (i.e. when she died), she had not had the event of interest. And as noted above, people who have not had the event when all follow-up ends for all subjects, are also censored. We can view this as a special type of censoring, because everyone who has not had the event or already been censored for some other reason, is censored at this time.

One of the appeals of survival analysis techniques is that we can include data (including information on covariates or independent variables of interest, such as treatment status) from subjects who are censored (either by drop-out, death, or some other competing event) up to the time that they are censored. For example, in this hypothetical study, if we were only recording whether or not a person had the event of interest during the full study period – i.e. our dependent variable was a dichotomous yes/no – then we might well have to completely drop cases #2 and 4 because we don't know whether or not they had an event during the full time window of the study. Additionally, of course, survival analysis allows us to examine not just whether an event occurred but how long it took to occur, which can also add considerable power to a study, particularly if the study is evaluating a treatment designed to delay (but possibly not prevent entirely) some undesired endpoint.

A BRIEF INTRO TO THE EXAMPLE DATA

The study from which the example data for this paper are drawn was a longitudinal observational study of the association between elder mistreatment and nursing home placement. Elder mistreatment includes physical or psychological abuse, as well as neglect by a responsible caregiver, and the study also evaluated 'self-neglect', the term for the situation where an older person in the community, is failing to adequately take care of him or herself. The research question was whether or not mistreated or self-neglecting older adults were more likely to be admitted to a nursing home – or be admitted to nursing homes sooner -- than older adults who were not identified as being mistreated or self-neglecting,

controlling for other factors that might increase risk of nursing home placement. The study population was a cohort of about 2,800 persons 65 and older living in New Haven, Connecticut who enrolled in a large study of aging in 1982. These persons were interviewed approximately every year for twelve years, from which we obtained data on a large number of risk factors for nursing home placement, such as social support, cognitive status and functional ability (e.g. ability to prepare meals, bath and dress oneself). To obtain information on elder mistreatment, nursing home placement and mortality, we conducted a record linkage to three other data sources: (1) Adult Protective Services records -- to determine if (and when) each person had been the victim of elder mistreatment or was identified as self-neglecting; (2) the Connecticut Long-term Care Registry -- to determine if (and when) each person had been admitted to a nursing home; and (3) death records to determine if and when the person had died. These records covered the time period of the study.

Thus, in this study, we have the timing of the outcome events, the timing of censorship, and indeed our main independent variable of interest changes over time (i.e. is time-dependent). Specifically, at baseline, none of the participants had been reported to protective services – those that were so reported during the study, thus became “exposed” at different times, which is a key feature of the analysis. Of course, for this paper, the purpose of which is mainly to teach about survival analysis using SAS, I have left out lots of study details and am not focusing on the findings; for more information about the real study, see (Lachs, Williams et al. 1997; Lachs, Williams et al. 1998; Lachs, Williams et al. 2002) and (Foley, Ostfeld et al. 1992).

FIRST STEP – CONSTRUCT SURVIVAL TIME AND CENSORING VARIABLES

Before we can do any survival analysis, we need to make sure that our data are structured appropriately and that we have constructed the needed variables for our outcome – which are the survival time variable and the censoring variable. We need to construct these variables for every case in the data set, whether or not the person has the event of interest or is censored. Let’s give a conceptual definition of each of these, before we dive into SAS code:

- **Survival time** – for an individual subject, time from study start (that is, when we started observing this person for an event) to when one of three things occurs. Note that if more than one of these things happens, we will choose the earliest. Also note that time can be measured in any units (e.g. days, months or even years – in some laboratory studies it might be hours or minutes), but for the methods described in this paper, it needs to be essentially continuous because it is very important that we be able to *order* events precisely, and if time is too crudely measured, there will be lots of tied survival times, which can cause problems.
 1. He/she has the event of interest
 2. He/she has some other event that makes him/her no longer at risk for the event – this could be dropping out of the study or having some other event (e.g. death) that precludes him/her from having the event of interest.
 3. The study ends – that is, we stop observing all study participants for event occurrence
- **Censoring indicator** – exactly how this is defined will differ from one study to another, but essentially we need to have a variable – defined for all participants – that allows us to distinguish whether a given individual's survival time represents time to the event of interest (i.e. #1 above) or time until some other competing event or end of study (i.e. #2 or 3).

In our example study, I am starting at the point where we have combined all our data sources, and we have a single record for each study participant. We will construct the above variables from several other variables that we have on our data set, namely

- **BASEDATE** – a SAS date variable (i.e. an integer that is the number of days elapsed from Jan 1, 1960 to the date of interest) indicating the date that the participant was enrolled in the study, i.e. the date of his/her baseline interview. In this study, these interviews were all conducted between February and December of 1982.
- **NHADMIT** – a 0,1 indicator of whether the person had a nursing home placement during study follow-up – between the baseline date and the end of the study (December 31, 1995).
- **NHPDATE** – a SAS date variable indicating the date when the participant was first admitted to a nursing home. This variable is missing if the person had not been admitted to a nursing home by the end of the study.
- **DIED** – a 0,1 indicator of whether the person died during study follow-up – between the baseline date and the end of the study (December 31, 1995).

- **DEATHDATE** – a SAS date variable indicating the date when the participant died. This variable is missing if the person had not died by the end of the study.

A couple of additional notes about these dates are important. First, note that none of them have anything to do with whether or when the person was identified as a victim of elder mistreatment or self-neglecting – this is appropriate, because our outcome definition should be independent of our risk factor/treatment definition. Of course, we will use information about elder mistreatment later, in defining those variables – as time-dependent covariates. Second, because – in this study – the ascertainment of our endpoints (nursing home placement and death) did not require continued study participation, we do not have any loss-to-follow-up; that is, our only sources of censoring are the end of the study follow-up or death. However, the methods described here are directly applicable if there are other reasons for censoring (ignoring statistical details such as that this censoring/study drop-out might be related to whether or not the person had the outcome of interest but we just don't know it).

Ok, given these source variables, let's define our survival time and censoring indicator variables, calling these EVENTDYS and CENSOR, respectively. Note that these are NOT special SAS variable names – we can call them anything we want – the way they are used in our programs will tell SAS that they are the survival time and censoring variables. EVENTDYS is the number of days from study start to the earliest of nursing home placement, death or end of study. CENSOR defines which of these three 'events' defines the end for each person: specifically, nursing home placement (CENSOR=0), death (CENSOR=1), or end of study (CENSOR=2). The following code, within a DATA step will define these variables:

```
endfwupdate = MDY(12,31,1995);
IF (nhadmit = 1) AND (basedate LE nhdate LE endfwupdate) THEN DO;
    censor = 0;
    censdate = nhdate ;
END;
ELSE IF (died = 1) AND (basedate LE deathdate LE endfwupdate) THEN DO;
    censor = 1;
    censdate = deathdate ;
END;
ELSE IF (died NE 1) OR (deathdate GT endfwupdate) then do;
    censor = 2;
    censdate = endfwupdate ;
END;

** time on study -- baseline to nh admit/death/end of study ;
eventdys = censdate - basedate ;

LABEL censor = 'Type of event'
       censdate = 'Date of NH/death/end fwp'
       eventdys = 'Days from baseline to NH/death/end fwp'
       ;
```

It was not essential to define the variable ENDFWUPDATE since it has the same value for all observations in this study, but it contributes to the clarity of the code and allows for the possibility that it might vary in other situations. Similarly, I could have done without CENSDATE by just defining EVENTDYS within each IF-THEN-DO block; I simply find the way I've done it here a little clearer.

NEXT STEP – EXAMINE SIMPLE SURVIVAL CURVES

Finally, some analysis! One of the first analyses of survival data is usually plotting some survival curves, and PROC LIFETEST has this covered. This first program just plots the survival distribution function, using the Kaplan-Meier (or product-limit) method.

```
PROC LIFETEST DATA = em_nh1 METHOD=KM PLOTS=SURVIVAL;
TIME eventdys*censor(1,2) ;
TITLE1 FONT="Arial 10pt" HEIGHT=1 BOLD 'Kaplan-Meier Curve -- overall';
RUN;
```

On the PROC LIFETEST statement, we specify that the method we want to use is the Kaplan-Meier method (METHOD=KM); it is also known as the product-limit method (METHOD=PL is synonymous),

and, in fact it is the default method, but I like to be explicit about such things. The alternative is the life-table or actuarial method (METHOD=LT, which is more suitable for very large data sets, and when measurement of event times is not precise); which I'm not covering here. I also indicate that we want to see the survival plot (PLOTS=SURVIVAL). In the TIME statement, using the syntax shown, we specify what our event time variable is (EVENTDYS), what the name of the censoring variable is (CENSOR), and, in the parentheses, the value (or values) of that variable that indicate that the observation is censored (here, as described above 1 & 2). We'll get to the graph itself momentarily, but first a look at some of the printed output (**Output 1**).

Output 1. Subset of Product Limit (aka Kaplan Meier) Estimates

The LIFETEST Procedure					
Product-Limit Survival Estimates					
EVENTDYS	Survival		Number	Failed	Left
	Standard	Failure			
0.00	1.0000	0	0	0	2769
2.00*	.	.	.	0	2768
6.00	0.9996	0.000361	0.000361	1	2767
7.00	.	.	.	2	2766
7.00	0.9989	0.00108	0.000625	3	2765
18.00*	.	.	.	3	2764
19.00*	.	.	.	3	2763
22.00*	.	.	.	3	2762
25.00	0.9986	0.00145	0.000722	4	2761
26.00*	.	.	.	4	2760
28.00*	.	.	.	4	2759
32.00	0.9982	0.00181	0.000808	5	2758
33.00*	.	.	.	5	2757
33.00*	.	.	.	5	2756
>>> SNIP <<<					
4974.00*	.	.	.	935	6
4975.00*	.	.	.	935	5
4975.00*	.	.	.	935	4
4977.00*	.	.	.	935	3
4977.00*	.	.	.	935	2
4978.00*	.	.	.	935	1
4978.00*	0.5256	0.4744	.	935	0
NOTE: The marked survival times are censored observations.					

What this table shows is a row for every value of EVENTDYS that has one or more events ("failures") or a censoring. (Note I have snipped out a large portion of the output – everything that happened between day 33 and day 4,974). If there are ties, multiple rows are shown (e.g. EVENTDYS = 33, EVENTDYS=4975). The EVENTDYS that have censored observations are shown with asterisks. The Survival column shows the proportion still surviving without an event (here without going into a nursing home) – it is only shown for EVENTS not censoring; thus, we see that an individual has a probability of 0.5256 of staying out of a nursing home for 4978 days (the entire study follow-up). Similarly, the number failed includes only events (NH placements), but the Number Left does go down with both events and censoring because it reflects the number remaining at risk for the event. Please see Paul Allison's wonderful book (in the reference list) for additional statistical details, on this and other aspects of survival analysis.

Output 2. Subset of Product Limit (aka Kaplan Meier) Estimates

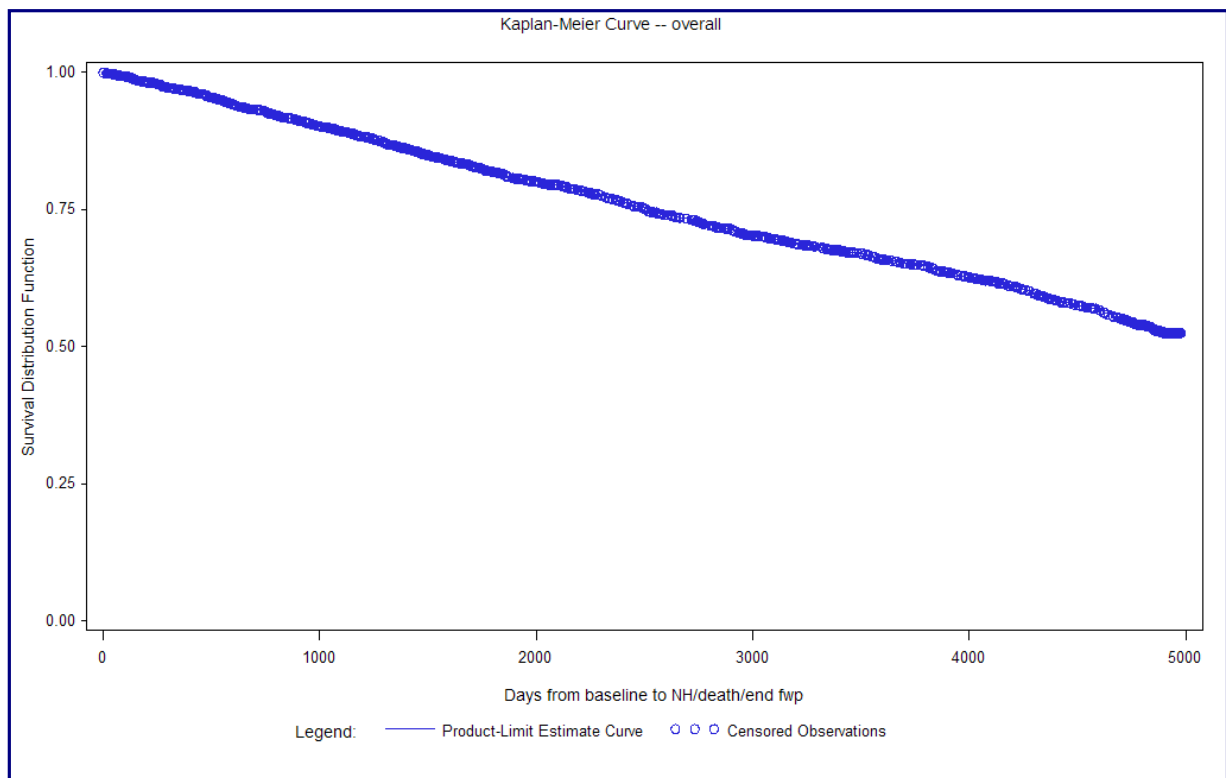
Summary Statistics for Time Variable EVENTDYS				
Quartile Estimates				
Percent	Point Estimate	95% Confidence Interval Transform	[Lower	Upper)
75	.	LOGLOG	.	.
50	.	LOGLOG	.	.
25	2512.00	LOGLOG	2373.00	2711.00
Summary of the Number of Censored and Uncensored Values				
	Total	Failed	Censored	Percent Censored
	2769	935	1834	66.23

This next chunk of output (**Output 2**) also gives some very useful summary information regarding the distribution of events over time. We see that 25% of participants had had an event by 2,512 days (about 6 years and 11 months); the study didn't last until the median survival time (i.e. fewer than half had been placed in a nursing home by the end of the study). As in the longer lifetable, it also shows that a total of 935 people had an event; the remainder were censored – either died during follow-up (without having entered a nursing home) or were alive and not in a nursing home when the study ended.

Now, moving onto the plots

In SAS 9.1 and later, the default for the plots is to turn on SAS Graph, and the output, without any extra tweaking of the program, is shown in **Output 3**. If you should want to not use GRAPH, specify LINEPRINTER on the PROC LIFETEST statement. PROC LIFETEST also takes advantage of ODS graphics in version 9.2, and I'll show a few options there later.

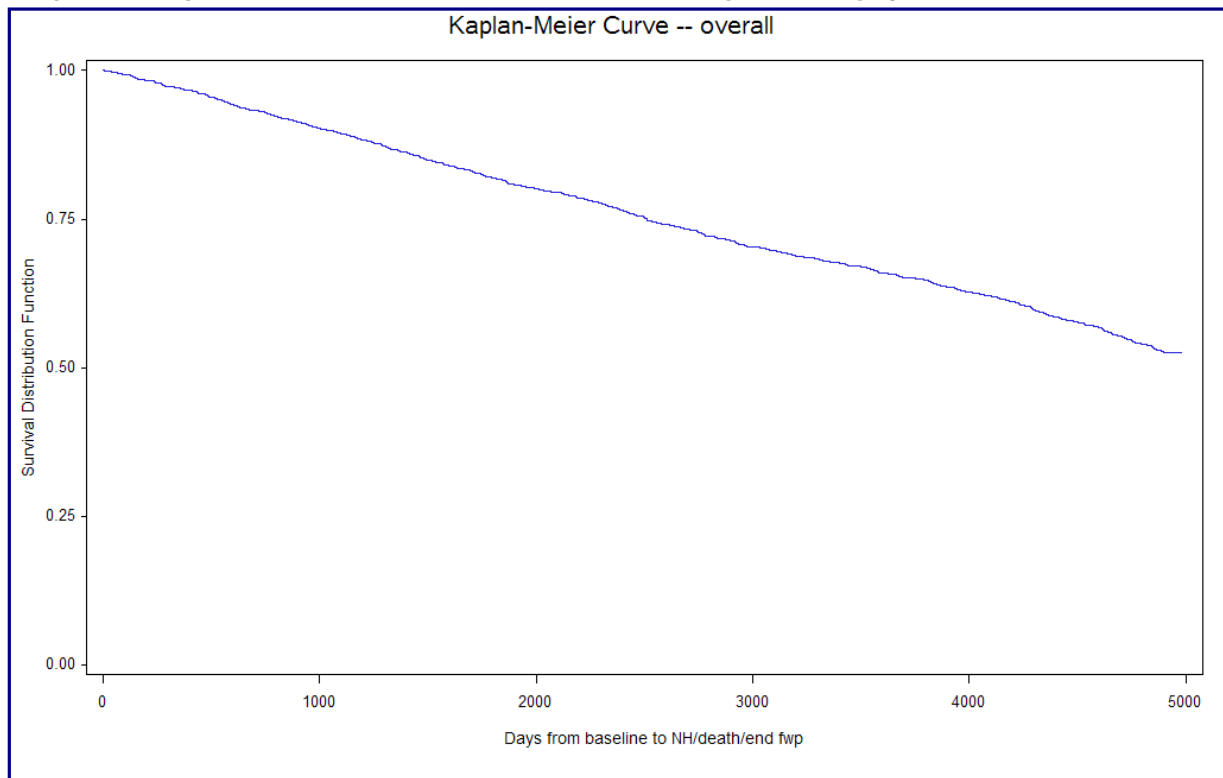
Output 3. Simple Plot of the Survivor Function



What is a little hard to tell from Output 3 is that it is basically a line of circles, which the legend tells us are the censored observations. In this study there are so many censorings, that it detracts from the curve. You can specify what symbol you want for the censored observations using the CENSORED SYMBOL= (or CS for short =) option on the PROC LIFETEST statement. I'm going to request no symbol – CS=none. The only other change in the program below is to specify the NOTABLE option so that it won't print the very long lifetable output again, and I'm using the S abbreviation for SURVIVAL in the PLOTS request. The resulting output is shown in **Output 4**.

```
PROC LIFETEST DATA = em_nh1 METHOD=KM PLOTS=S CS=NONE NOTABLE;
TIME eventdys*censor(1,2) ;
TITLE1 FONT="Arial 10pt" HEIGHT=1 BOLD 'Kaplan-Meier Curve -- overall';
RUN;
```

Output 4. Simple Plot of the Survivor Function, eliminating censoring symbol



TESTING FOR DIFFERENCES IN SURVIVAL CURVES

Frequently – both in experimental and observational research, one is interested in whether the “survival” of subjects differs based on covariates of interest, whether that is a treatment or some observed or measured characteristic. For a categorical variable, one can assess differences in survival using the STRATA statement in PROC LIFETEST. Below I show the example for marital status. Of course, this could be a time-dependent covariate, but for now, we are just considering the participant’s marital status at the time of study enrollment, coded 0 if unmarried, 1 if married. I also use a couple of SYMBOL statements (these are really ‘talking to SAS Graph’) so that, whether I used a color printer or not, I could tell the lines apart.

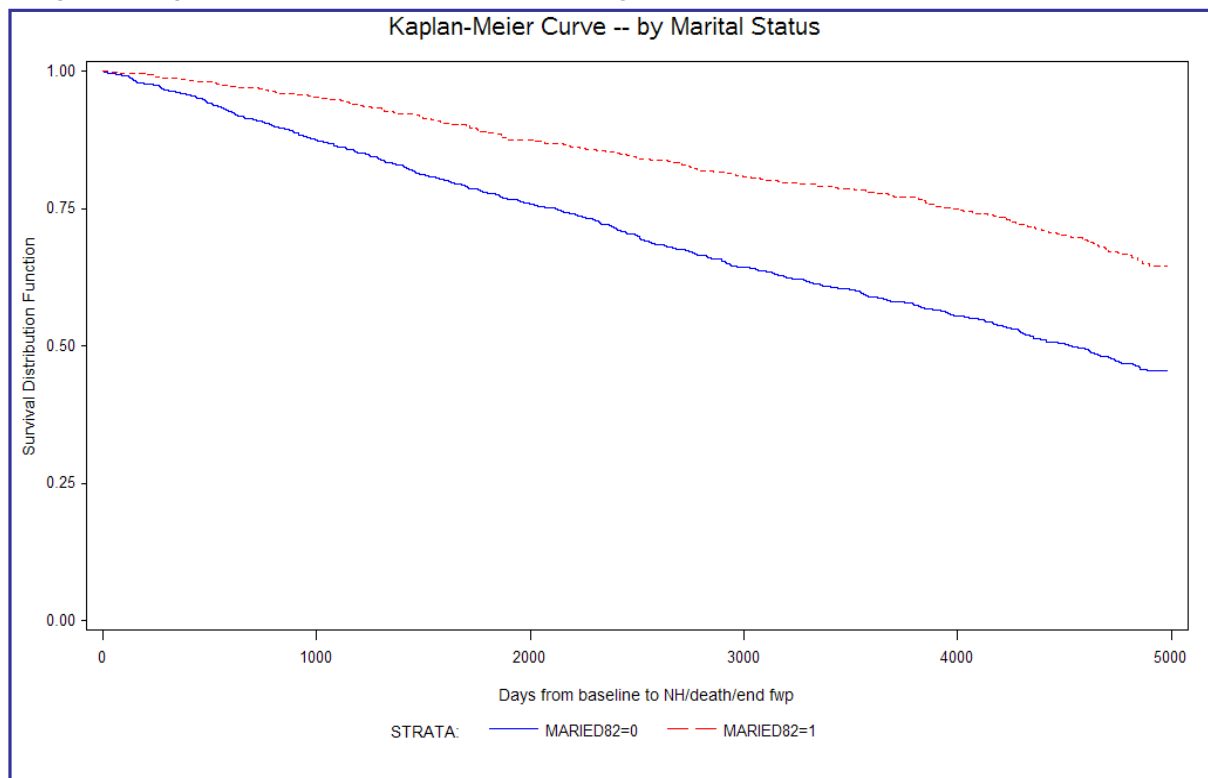
```
PROC LIFETEST DATA = em_nh1 METHOD=KM PLOTS=S CS=NONE;
TIME eventdys*censor(1,2) ;
STRATA married82 ;
SYMBOL1 V=none COLOR=blue LINE=1;
SYMBOL2 V=none COLOR=red LINE=2;
RUN;
```

A portion of the tabular output is shown in **Output 5**, and the survival plots are shown in **Output 6**.

Output 5. Testing for differences in survival curves for single dichotomous variable

Testing Homogeneity of Survival Curves for EVENTDYS over Strata			
Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	83.8522	1	<.0001
Wilcoxon	86.8443	1	<.0001
-2Log(LR)	84.0352	1	<.0001
NOTE: 26 observations with invalid time, censoring, or strata values were deleted.			

Output 6. Kaplan-Meier survival curves according to marital status



In **Output 5**, We get three different statistical tests for whether married and unmarried people differ in their risk of nursing home placement. Clearly, there is a strong effect, but how do the three statistics? Without going into a lot of details, the log-rank test is most similar to what is tested in a proportional hazards model, and is most sensitive to differences in survival later in follow-up, while the Wilcoxon is more powerful at detecting differences earlier in followup. Either of these tests would be appropriate here. The likelihood ratio test (shown by -2 Log(LR)) is based on the assumption that the event times follow an exponential distribution, which is rarely met in health research; so I ignore this test. We also get a somewhat informative note that some observations have been deleted; in this case I know that there were a few people for whom marital status was unknown – always good to check that such notes agree with your expectations regarding your data set!

With respect to the curves (**Output 6**), of course, the p-values had clued us in that married and unmarried participants were at significantly different risk of nursing home admission. The curves confirm what would likely have been our intuition – that the unmarried (usually in this age group – this means widowed) were much more likely to go into a nursing home than those who were married. Notably the curves are also fairly parallel, suggesting perhaps that the hazards are proportional.

STRATIFYING ON MULTIPLE FACTORS AND TESTING FOR DIFFERENCES FOR CONTINUOUS VARIABLES

You may wish to test for differences in event-free survival for multiple covariates, and some may be quantitative or continuous variables. There is a limit to how much of this is practical without moving onto regression analysis, but LIFETEST does have some capabilities in this regard. Multiple variables can be placed on the STRATA statement, and the TEST statement produces a test for differences for one or more continuous variables. In the example below, I'm putting two variables on the STRATA statement and three on the TEST statement, so we can see what happens. I've added gender as a STRATA variable, and baseline age, body mass index and depression score as TEST variables; otherwise the code is the same.

```

PROC LIFETEST DATA = em_nh1 method=km plots=(survival) cs=none NOTABLE;
TIME eventdys*censor(1,2) ;
STRATA married82 gender;
TEST age82 bmi82 cesd82 ;
SYMBOL1 V=none COLOR=red LINE=1 width=2;
SYMBOL2 V=none COLOR=blue LINE=1 width=2;
SYMBOL3 V=none COLOR=red LINE=2 width=2;
SYMBOL4 V=none COLOR=blue LINE=2 width=2;
RUN;

```

A portion of the tabular output is shown in **Output 7** (for the STRATA variables) and **Output 9** (for the TEST variables); the survival curves are shown in **Output 8**.

Output 7. Including multiple STRATA and TEST variables – STRATA results

Summary of the Number of Censored and Uncensored Values

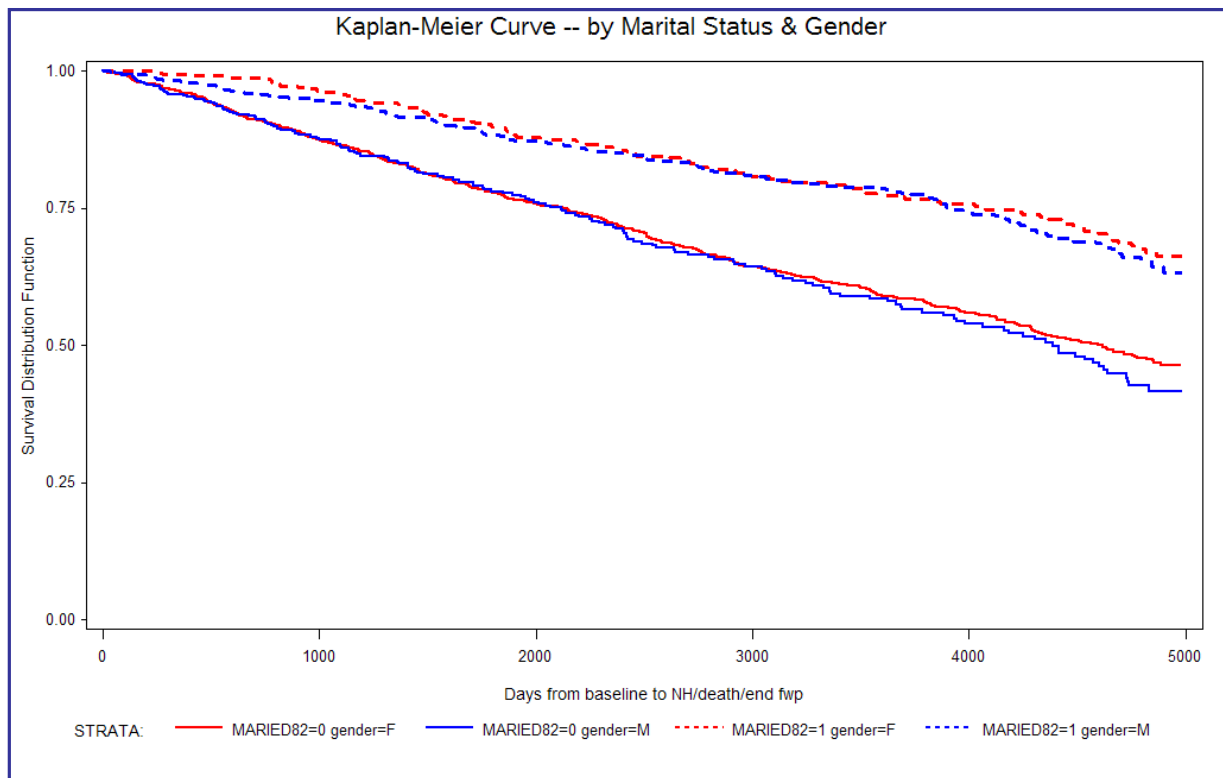
Percent Stratum	MARIED82		gender	Total	Failed	Censored	Censored
1	0	F		1236	516	720	58.25
2	0	M		471	164	307	65.18
3	1	F		367	94	273	74.39
4	1	M		669	150	519	77.58
-----				-----			
Total				2743	924	1819	66.31

Test of Equality over Strata

Pr > Test	Chi-Square	DF	Chi-Square
Log-Rank	84.9756	3	<.0001
Wilcoxon	87.4265	3	<.0001
-2Log(LR)	84.4140	3	<.0001

The first thing to note about the output for the STRATA variables (**Output 7**), is that it is truly a stratified analysis – the data have been broken into four categories (married men, married women, unmarried men and unmarried women) and is testing whether there are any differences among these four strata (hence the tests all have three degrees of freedom (DF = 3) we are not getting separate tests of marital status, controlling for gender or gender controlling for marital status, nor are we getting univariate (unadjusted) tests for either of these variables. This is also evidenced by the fact that the survival plot (**Output 8**) has 4 separate curves. I'll come back in a minute to how one would get those, but in the meantime we see that there are clear differences among the three strata, though we are not testing any specific hypotheses about either variable individually – these are exactly the same results that we would get if we had created a single variable with four categories for the possible combinations of gender and marital status. Note that this also shows why it becomes impractical to include LOTS of variables on your STRATA statement, because the number of categories becomes big very quickly...if we had just seven dichotomous variables, there would be 128 strata (2^7), which – even if the study was large enough to support such an analysis, would be virtually uninterpretable! If we wanted to get separate unadjusted tests for a number of categorical covariates, it is perfectly fine in PROC LIFETEST to include multiple STRATA statements – then the strata formed by whatever variables are on each STRATA statement are tested separately.

Output 8. Including multiple STRATA and TEST variables – Kaplan Meier Curves



We see in the survival curves that within categories of marital status, the survival curves (time to nursing home placement) are quite similar for men and women. As noted above we can separately test the effects of marital status and gender by including two STRATA statements. As it turns out, gender is significant unadjusted for marital status (results not shown here). As we'll see below with the explanation of the TEST statement, we could test whether gender has an effect controlling for marital status (and vice versa), by including both on a TEST statement. While the TEST statement is designed for quantitative variables, by coding gender as numeric (e.g. 1 for male and 0 for female), we could use TEST to evaluate the incremental effect of gender, controlling for marital status – or we could include one on the TEST statement and one on the STRATA statement.

Output 9. Including multiple STRATA and TEST variables – TEST results

Rank Tests for the Association of EVENTDYS with Covariates Pooled over Strata						
Univariate Chi-Squares for the Log-Rank Test						
Variable	Test Statistic	Standard Error	Chi-Square	Pr > Chi-Square	Label	
AGE82	-3170.9	158.5	400.3	<.0001	Baseline AGE	
BMI82	747.8	133.9	31.1807	<.0001	1982-Body Mass Index	
CESD82	-1607.5	222.3	52.3079	<.0001	CESD82: Depression	
Forward Stepwise Sequence of Chi-Squares for the Log-Rank Test						
Variable	DF	Chi-Square	Pr > Chi-Square	Chi-Square Increment	Pr > Increment	Label
AGE82	1	400.3	<.0001	400.3	<.0001	Baseline AGE
CESD82	2	450.6	<.0001	50.2938	<.0001	CESD82: Depression
BMI82	3	460.1	<.0001	9.4513	0.0021	1982-Body Mass Index

The TEST statement also produces both Log-Rank and Wilcoxon results; however, they are virtually identical, so I'm including just the Log-Rank tests (**Output 9**). We also get two sets of results with respect

to the hypothesis being tested – Univariate and Forward Stepwise. The Univariate results are testing each variable singly; that is, not adjusted for the other; all are highly statistically significant. The signs of the log-rank test statistic tell us the direction of the results. For example, the negative sign for AGE82 indicates that those who were older at baseline have shorter times to nursing home placement; the same is also true for those with higher depression scores. In contrast, those with a greater body mass index have longer times to nursing home placement (in this age group, greater weight usually means less frailty).

In the Forward Stepwise results, LIFETEST first finds the variable with the highest chi-square, and includes it in the set to be tested. This variable is AGE82, and since it is the first variable in (and just coincidentally the first one listed on the TEST statement), the results for this are identical in the two sets of results. The second strongest variable is then added, and it is tested, controlling for the prior variable – so in the lower panel, the effect for CESD82 adjusts for AGE82 – its chi-square is slightly smaller than in the Univariate results but still highly significant. Note also that there are two chi-square statistics and tests in each row. The first one (value 450.6 in the CESD82 row) tests the null hypothesis that both the included variables are non-significant, while the one labeled ‘Chi-Square Increment’ (50.2938 in the CESD82 row) tests for the additional effect of that variable – it is the difference between the joint chi-square of the current and the prior row (i.e. $450.6 - 400.3 = 50.3$ – ignoring rounding error). Finally, BMI82 is added to the mix, and its effect is adjusted for both AGE82 and CESD82.

How the STRATA statement and the TEST statement affect each other

We’ve seen that TEST and STRATA are different in the way that they ‘control’ for other variables on the same statement, but to this point we haven’t mentioned how the presence of both the TEST and STRATA statements affect each other – and this is a bit tricky, as the effects are not the same in both directions. First, including a TEST statement along with one or more STRATA statements has no impact at all on the results of the STRATA statement(s) – that is, neither the survival curves nor the statistical tests produced by STRATA are in any way affected by the presence of the TEST statement. You can readily prove this to yourself by not changing anything in your program but the presence of the TEST statement and see that the STRATA results are identical. However, the converse is NOT true, and the first clue to this is the heading (see Output 8) in the TEST results that states “Rank Tests for the Association of EVENTDYS with Covariates Pooled over Strata”. What does “Pooled over Strata” mean? When there is a STRATA statement, both the Log-Rank and the Wilcoxon test statistics are first calculated within the categories formed by the STRATA statement, and then averaged (or pooled) across strata; in other words, they control for the STRATA variables.

This bears repeating: the results of the STRATA statement(s) are NOT affected by the presence of the TEST statement – the STRATA variables are not adjusted for the TEST variables. However, the results of the TEST statement are affected by the presence of the STRATA statement – the TEST variables are adjusted for the STRATA variables. This fact can be creatively exploited to test particular hypotheses. For example, to get a test of gender adjusted for marital status, we could include MARIED82 on the STRATA statement and gender (coded as Male = 1, Female = 0) on the TEST statement. To test for an effect of marital status, adjusted for gender, we could do the converse...Of course, we could also move on to modeling...

COX MODELS OR PROPORTIONAL HAZARDS REGRESSION

While I am not going to elaborate any statistical details and there are MANY features and applications of PHREG that I am not going to cover, an important benefit of the Cox model is that it doesn’t make any assumptions about the distribution of the event times and can readily accommodate both discrete (interval) and continuous event times. And, in fact, as Paul Allison persuasively argues, even though the model is called proportional hazards (and we’ll show a way to test that ‘assumption’), the importance of violations to this assumption have likely been overrated, and, more importantly, the model can readily be adapted to deal with non-proportionality. Indeed, because what proportionality essentially means is that the relative hazard (i.e. the effect of a covariate) is constant over time, time-dependent covariates allow for a specific type of non-proportionality.

First, let’s just show the syntax and results for a simple model with no time-dependent covariates. See below. Note that the two models shown below are virtually identical, but the second one will work only in SAS 9.2 – yes, PHREG finally has a CLASS statement! Most of the standard output is shown in **Output 10**.

```

PROC PHREG DATA = em_nh1 ;
MODEL eventdys*censor(1,2) = male age82 married82 bmi82 cesd82 /RL TIES=efron;
RUN;

* NEW IN 9.2 - CLASS STATEMENT!! ;
PROC PHREG DATA = em_nh1 ;
CLASS gender ;
MODEL eventdys*censor(1,2) = gender age82 married82 bmi82 cesd82 /RL;
RUN;

```

Output 10. PHREG output

Model Information									
Data Set			WORK.EM_NH1						
Dependent Variable			EVENTDYS		Days from baseline to NH/death/end fwp				
Censoring Variable			CENSOR		Type of event				
Censoring Value(s)			1 2						
Ties Handling			EFRON						
Number of Observations Read					2769				
Number of Observations Used					2436				
Class Level Information									
Class			Value		Design Variables				
gender			F		1				
			M		0				
Summary of the Number of Event and Censored Values									
Total			Event		Censored		Percent Censored		
2436			782		1654		67.90		
Testing Global Null Hypothesis: BETA=0									
Test			Chi-Square			DF		Pr > ChiSq	
Likelihood Ratio			465.6779			5		<.0001	
Score			545.7160			5		<.0001	
Wald			526.5441			5		<.0001	
Type 3 Tests									
Effect			DF		Wald Chi-Square		Pr > ChiSq		
gender			1		5.5483		0.0185		
AGE82			1		347.7227		<.0001		
MARRIED82			1		23.0021		<.0001		
BMI82			1		10.5311		0.0012		
CESD82			1		47.7229		<.0001		
Analysis of Maximum Likelihood Estimates									
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
gender	F	1	-0.18902	0.08025	5.5483	0.0185	0.828	0.707	0.969
AGE82		1	0.09663	0.00518	347.7227	<.0001	1.101	1.090	1.113
MARRIED82		1	-0.41698	0.08694	23.0021	<.0001	0.659	0.556	0.781
BMI82		1	-0.02727	0.00840	10.5311	0.0012	0.973	0.957	0.989
CESD82		1	0.02573	0.00372	47.7229	<.0001	1.026	1.019	1.034

The MODEL statement is a bit of a mix between what we saw in PROC LIFETEST (the event time and censoring are specified identically as in LIFETEST), while the independent variables are specified much

like in many other SAS regression procedures. We see at the top of the output that the Ties Handling method is Efron (which I specified with TIES=EFRON on the MODEL statement. The default method is BRESLOW, which is computationally less intense, but doesn't work so well when there are lots of ties. You can also specify TIES=EXACT, which is the most accurate, but also the most computationally demanding. My practical advice is to try all three with some simple models. If the results don't change very much, choose one of the faster methods for building your model.

The next part of the output, "Testing Global Null Hypothesis," is simply providing three different tests of the hypothesis that ALL of the parameter estimates for the included covariates are 0 (i.e. nothing in the model is statistically significant). In my experience these usually are quite similar in what they indicate – and if it is non-significant you need to find a new model!

The TYPE3 Wald Tests are somewhat redundant here with the Maximum Likelihood estimates, but if any of our variables had more than 1 degree of freedom (e.g. a multi-category CLASS variable), this table would give an overall test for that variable – adjusting for all other covariates. This is a useful, new feature of PHREG in conjunction with the CLASS statement.

The RL option (short for RISKLIMIT) gets SAS to compute and print the hazard ratio and its 95% confidence limit; of course the hazard ratio is simply the exponentiation of the parameter estimate. When the parameter estimate is negative the hazard ratio will be less than one (indicating that the variable is associated with decreased risk or longer times-to-event); hence, as we saw in our LIFETEST analyses, women, those who are married and those with higher BMI are at lower risk of nursing home placement; while those who are older and have higher depression scores are at increased risk – note that the estimates are per unit increase – so the hazard increases by 10% for each additional year in age, and by about 2% for each additional point on the CESD depression score (which, by the way has a range of 0-60). All of the included effects are statistically significant.

CODING AND TESTING TIME-DEPENDENT COVARIATES

In my mind, this is one of the coolest aspects of Cox regression, and in particular, the way it is implemented in SAS. Several of the regression procedures now allow you to include DATA STEP-like code for specifying covariates...but what is unusual – and tricky! about this code in PHREG is that the time-dependent covariates that are created by DATA STEP-like statements in PHREG could NOT be created in the DATA STEP. Figure 2 attempts to give a conceptual understanding of why this is.

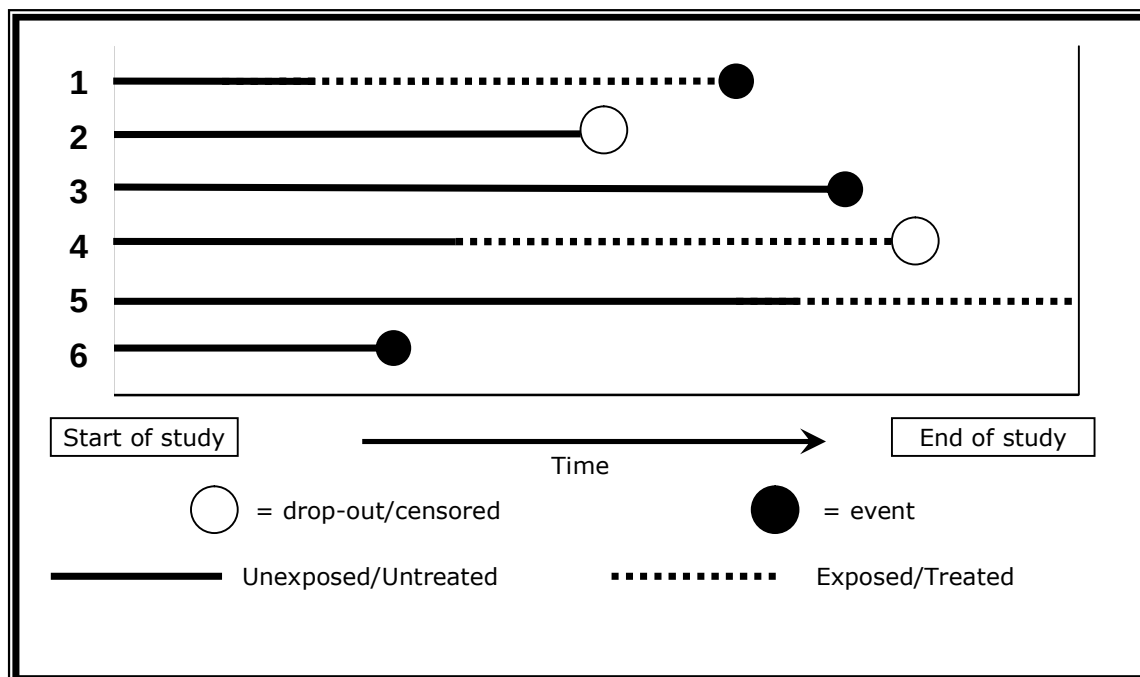


Figure 2. Schematic depiction of survival data for six patients, with a time dependent covariate. See text for further description.

What is different about this figure from Figure 1 back at the beginning of this paper is that I've introduced a covariate; further, this covariate is *time-dependent*; that is, its value changes over time. There are many different ways that such changes can be assessed (e.g. at regular intervals or continuous ascertainment), but in this example, individuals can go from being "unexposed" to the covariate or treatment of interest (all start out that way) to "exposed" – starting treatment, becoming positive for a given risk factor at different times during follow-up. In the study being used in this paper, we had exact dates when elder mistreatment or self-neglect were reported, and so (leaving aside delays in reporting), we know – to the day – when individuals convert from unexposed to exposed.

The way that the Cox model works is that it only "cares" about the exposure status of the population when an event occurs. In particular, in this example, the relative hazard estimate is based on the covariate profile of the population at risk (those not yet censored) when someone in the sample goes into the nursing home. This is the key to thinking about how time-dependent covariates are coded. When an event occurs in the sample (i.e. at a given value of EVENTDYS as we march through study time), we can think of PHREG as evaluating the relative hazard based on covariate values on that day. For example, in Figure 2, when #6 has his/her event, he/she was not exposed, but further among those still at risk (i.e. #1 - #5), only one was exposed (#1); the rest were unexposed. In contrast, when #1 has an event, only 4 people are still at risk, and two out of four are exposed. As long as we can specify what the exposure profile of the sample is at each event time, then the Cox model works – and this is what to think about in coding the time-dependent covariate in PROC PHREG.

This is shown in the code below, where we have two time-dependent covariates – whether or not someone has had a verified case of elder mistreatment (VEMS) or a verified case of self-neglect (VSN). Note that in the event that a person has had both, we are giving priority to elder mistreatment (viewed as more serious). Note also that these variables do NOT exist on the input data set (or if they did their values could change for the purposes of model estimation).

```
PROC PHREG DATA = em_nh1 ;
CLASS GENDER ;
MODEL eventdys*censor(1,2) = vems vsn
                                gender age82 married82 bmi82 cesd82 /RL TIES=EFRON;
IF (0 LE vemsdays LE eventdys) THEN DO;
    vems = 1;
    vsn = 0;
END;
ELSE vems = 0;
IF vems NE 1 THEN DO;
    IF (0 LE vslfdays LE eventdys) THEN vsn = 1;
    ELSE vsn = 0;
END;
RUN;
```

In contrast to VEMS and VSN (the time dependent covariates), the variables VEMSDYS and VSNDYS, which, along with EVENTDYS, contribute to the definition of VEMS and VSN do exist on the input data set. Specifically, VEMSDYS is the number of days (from baseline) when a person had a first verified complaint of elder mistreatment; similarly VSLFDYS is the days from baseline to an individual's first verified case of self-neglect. These time variables are static. However, the values of VEMS and VSN for each individual depend on the current value of EVENTDYS – not that individual's EVENTDYS but the EVENTDYS value at each non-censored EVENT-time. So, as time marches forward, every time there is an event for anyone in the sample, PHREG determines what each person's value of VEMS and VSN is, and the relative distribution of exposure status relative to event times is what determines the relative hazard for that variable. If we constructed these variables in the data step, they would depend only on the row or person-level value of EVENTDYS compared to that individual's timing of elder mistreatment or neglect – and would not change over time. When we code them in PHREG, they get constructed by comparing an individual's VEMSDYS and VSLFDYS values to the EVENTDYS value for every (non-censored) EVENTDYS in the sample. Note that – at least as of SAS 9.2 – time-dependent variables cannot be listed on the CLASS statement – thus we create two mutually exclusive 'dummy' variables for VEMS and VSN.

The output doesn't really indicate the sophisticated way in which the covariates have been defined...I'm showing only a portion of it here (**Output 11**). However, it shows that elder mistreatment, and particularly self-neglect greatly increase the risk of (shorten the time to) nursing home placement.

Output 11. PHREG output with a time-dependent covariate

Analysis of Maximum Likelihood Estimates								
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Confidence	Ratio Limits
vems	1	1.17229	0.23621	24.6301	<.0001	3.229	2.033	5.131
vsn	1	1.82335	0.13393	185.3456	<.0001	6.193	4.763	8.051
gender	F 1	-0.15470	0.08021	3.7200	0.0538	0.857	0.732	1.003
AGE82	1	0.09040	0.00511	313.4993	<.0001	1.095	1.084	1.106
MARIED82	1	-0.32184	0.08775	13.4519	0.0002	0.725	0.610	0.861
BMI82	1	-0.02272	0.00850	7.1513	0.0075	0.978	0.961	0.994
CESD82	1	0.02229	0.00379	34.6301	<.0001	1.023	1.015	1.030

Note that a similar strategy for coding could be used if covariates only changed at set times – say when a re-assessment occurred. For example, if we had interview dates IDATE82 – IDATE90 and covariate values CESD82 – CESD90, MARIED82 – MARIED90, etc, corresponding to the values of those covariates assessed at each interview, we could update these covariates (constructing AGE CESD, MARIED, etc) with code like the following:

```
PROC PHREG DATA = em_nh1 ;
MODEL eventdys*censor(1,2) = age cesd married /RL TIES=EFRON;
ARRAY idate{9} idate82-idate90;
ARRAY agey{9} age82-age90 ;
ARRAY dep{9} cesd82-cesd90 ;
ARRAY mar{9} married82-married90;
DO i = 1 TO 8;
  IF idate{i} LE (basedate + eventdys) < idate{i+1} THEN DO;
    age=agey{i} ;
    cesd = cesd{i} ;
    married = married{i} ;
  END;
END;
RUN;
```

This code may not be terribly efficient, as it requires processing the arrays every time there is an event, but it gets the job done, by assigning values to the covariates based upon the most recent assessment prior to the current EVENTDYS value.

Other types of situations and coding methods arise – the key thing to remember is that you need to be able to specify what each person's covariate values should be every time *anyone* in the sample has an event.

TESTING PROPORTIONALITY

As noted in the beginning of the section on Cox models, the problem of proportionality may be over-rated – perhaps because of the name that has been attributed to Cox models, namely *proportional* hazards models. Nonetheless, when viewed as a change in the relative hazard over time, one can think of the lack of proportionality as an interaction of time and a covariate...this not only suggests a way of testing it, but also can suggest remedies. Specifically, a particular type of time-dependent covariate is an interaction between EVENT-time and a covariate. For example, if we want to test whether the effect of gender varies over time, we can use this code.

```

PROC PHREG DATA = em_nh1 ;
MODEL eventdys*censor(1,2) = age82 cesd82 married82 male male_time /RL TIES=EFRON ;
male_time = eventdys*male ;
RUN;

```

A portion of the output is shown in **Output 12**.

Output 12. One method of testing proportionality

Analysis of Maximum Likelihood Estimates								
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
AGE82	1	0.10036	0.00472	452.8266	<.0001	1.106	1.095	1.116
CESD82	1	0.02807	0.00341	67.8487	<.0001	1.028	1.022	1.035
MARIED82	1	-0.43747	0.08264	28.0205	<.0001	0.646	0.549	0.759
MALE	1	-0.03274	0.13364	0.0600	0.8065	0.968	0.745	1.258
male_time	1	0.0001046	0.0000509	4.2158	0.0400	1.000	1.000	1.000

These results suggest that there is a minimal amount of un-proportionality for gender. While this makes the coefficient for MALE harder to interpret, the model itself has accounted for the lack of proportionality by adjusting for it. Because hazards are estimated on a log-scale, some recommend that the interaction be constructed with LOG(eventdys). If proportionality is a concern, it is probably worth testing it both ways. My advice also is to examine the stratified survival curves (including the Log-survival curves) to see if there is a particular time point at which the relative hazard seems to change, and consider building that into the model to aid in interpretation.

SAS 9.2 also offers another way of testing proportionality, which is the ASSESS PROPORTIONALITY statement, which can simultaneously test for proportionality of all the variables in the model. The statistical methods underlying this test are quite complex, however, so I encourage you to read the SAS documentation in order to use it correctly and interpret the results; see also Gharibvand & Fernandez (2008).

CONCLUSIONS

My goal in this paper has been to give an applied, not terribly technical introduction to some of the most widely used methods in survival analysis. In particular, I have tried to give an intuitive understanding of how this method works and gone into substantial detail into some of the simpler but very powerful methods of examining event-time data. There's a lot more to explore, including great new graphical capabilities with ODS graphics in SAS 9.2, and other sophisticated statistical features of PHREG. In particular, check out the new HAZARDRATIO statement, which I haven't covered but which offers a lot of flexibility in estimating hazard ratios for CLASS variables – even in the case of interactions. Nonetheless, I hope that this paper has taken a little of the fear away from these methods and given you some new tools to apply. Paul Allison's book is an invaluable companion for anyone doing survival analysis – I just wish a new edition would be published as many features of these PROCs have been enhanced since the first edition was published in 1995.

ACKNOWLEDGMENTS

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.

REFERENCES

- Allison, Paul D., *Survival Analysis Using the SAS® System: A Practical Guide*, Cary, NC: SAS Institute Inc., 1995. 292 pp.
- Foley, D. J., A. M. Ostfeld, et al. (1992). "The risk of nursing home admission in three communities." *J Aging Health* 4(2): 155-73.
- Gharibvand, L., Fernandez, G. (2008) "Advanced Statistical and Graphical features of SAS® PHREG" SAS Global Forum 2008 Proceedings <http://www2.sas.com/proceedings/forum2008/375-2008.pdf>

Lachs, M. S., C. Williams, et al. (1997). "Risk factors for reported elder abuse and neglect: a nine-year observational cohort study." Gerontologist **37**(4): 469-74.

Lachs, M. S., C. S. Williams, et al. (2002). "Adult protective service use and nursing home placement." Gerontologist **42**(6): 734-9.

Lachs, M. S., C. S. Williams, et al. (1998). "The mortality of elder mistreatment." JAMA **280**(5): 428-32.

SAS Institute Inc. *SAS/STAT 9.2 Users' Guide. Chapter 64: The PHREG Procedure* Cary, NC: SAS Institute Inc.

SAS Institute Inc. *SAS/STAT 9.2 Users' Guide. Chapter 49: The LIFETEST Procedure* Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

I welcome comments, suggestions and questions at:

Christianna S. Williams, PhD
Christianna_Williams@abtassoc.com