

Survival Models in SAS - Part 1: PROC LIFETEST

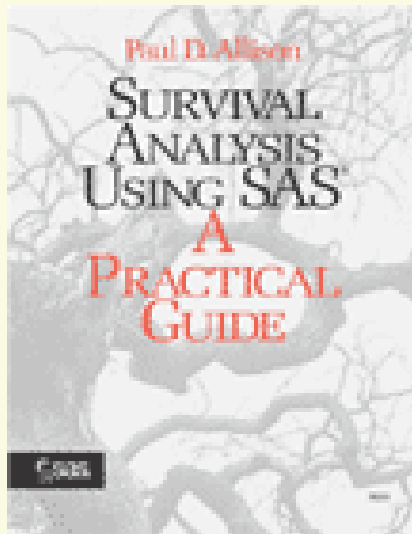
October 24, 2007

Charlie Hallahan

Chapter 1: Introduction

These talks are based on the book “**Survival Analysis Using the SAS System: A Practical Guide**” (1995) by Paul Allison.

The book is part of the SAS Books-by-Users series and can be found at <http://www.sas.com/apps/pubscat/bookdetails.jsp?catid=1&pc=55233>



Chapter 1: Introduction

This series of talks will cover

Chapter 1: Introduction

Chapter 2: Basic Concepts of Survival Analysis

Chapter 3: Estimating and Comparing Survival Curves with PROC LIFETEST

Later presentations will cover

Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG

Chapter 5: Estimating Cox Regression Models with PROC PHREG

Chapter 1: Introduction

As the author states: “*Survival analysis* is a class of statistical methods for studying the occurrence and timing of events”.

Since early applications involved the event being **death**, the name *survival analysis* appears to narrow the range of applications to which the methods of survival analysis could be applied.

In general, any time the random variable of interest, T , is the time until some event occurs, then the methods of survival analysis apply.

Other names for survival analysis are event history analysis, duration analysis, transition analysis, failure time analysis, and reliability analysis.

Chapter 1: Introduction

Two common features of survival analysis are:

- **censoring**: for some observations, the event may not have occurred yet or has already occurred by the start of the study
- **time-dependent covariates**: explanatory variables that change over time

Many different methods can be applied to survival data:

- life tables
- Kaplan-Meier estimators
- exponential regression
- log-normal regression
- proportional hazards regression
- competing risks models
- discrete-time methods

Chapter 1: Introduction

SAS/STAT software has three main procedures that can be used for survival analysis:

LIFETEST: designed for univariate analysis of the timing of events

LIFEREG: estimates regression models with censored, continuous-time data under several distributional assumptions

PHREG: uses Cox's partial likelihood method to estimate regression models with censored data. It allows for time-dependent covariates and handles both continuous-time and discrete-time data.

Chapter 2: Basic Concepts of Survival Analysis

This chapter covers several topics common to many methods used in survival analysis:

- **censoring**: a nearly universal feature
- **survivor & hazard functions**: ways to represent the probability distribution of time-to-event data
- **choice of origin**: when do measurements start?
- **basic data structure**: how to organize the data for analysis?

Chapter 2: Basic Concepts of Survival Analysis

Censoring

Most basic distinction is between *left censoring* and *right censoring*.

An observation is *right censored* at some value c if all we know is that $T > c$.

An observation is *left censored* at some value c if all we know is that $T < c$. (i.e., the event of interest happened before we collected the data).

An observation is *interval censored* if all we know is that $a < T < b$.

LIFEREG can handle all three kinds of censoring.

Chapter 2: Basic Concepts of Survival Analysis

A distinction is made between several kinds of right censoring.

An observation is said to be *singly Type I censored* if the censoring time is fixed (*Type I*) and all observations have the same censoring time (*singly*).

For example, data is collected for a fixed period of time and some observations have not had the event of interest occur by the end of the data collection.

Type II censoring occurs when an observation is terminated after a prespecified number of events have occurred.

For example, an experiment with 100 rats may be stopped after 50 have died. This type of censoring is not as common in the social sciences.

Chapter 2: Basic Concepts of Survival Analysis

Random censoring occurs when the reason that observations are censored is not under the control of the analyst.

For example, subjects leave the study for no known reason or enter the study at random times (eg., a study of survival after an operation when the date of the operation varies across subjects).

Standard methods of survival analysis do not distinguish among Type I, Type II, and random censoring. They are all treated as generic right-censoring.

The likelihood methods discussed have no problem with Type I and Type II censoring, but any random censoring must be assumed to be *noninformative*.

Cox and Oates (1984): “A crucial condition is that, conditionally on the values of any explanatory variables, the prognosis for any individual who has survived to c_i should not be affected if the individual is censored at c_i . That is, an individual who is censored at c should be representative of all those subjects with the same values of the explanatory variables who survive to c ”

Chapter 2: Basic Concepts of Survival Analysis

An example of *informative random censoring* would occur in a study of how long people stay unemployed and some subjects just drop out of the job market completely. These dropouts are most likely people who would have stayed unemployed longer than those who remained in the study and kept looking for work.

In principal, informative censoring can lead to severe biases.

There is no statistical test for informative versus noninformative censoring.

Chapter 2: Basic Concepts of Survival Analysis

Describing Survival Distributions

Let T be a non-negative random variable representing the time to an event.

Survival analysis deals with a number of functions associated with T .

$f(t)$ = the **probability density function** for T

$F(t)$ = the **probability distribution function** for T ; $F(t) = \text{Prob}(T \leq t)$

Instead of these two common functions, survival analysis concentrates on two related functions:

$S(t) = 1 - F(t) = \text{Prob}(T > t)$ = probability of **surviving beyond time t** .

$S(t)$ is called the **survivor function**.

Related to the survivor function is the fundamental **hazard function**.

Chapter 2: Basic Concepts of Survival Analysis

The hazard function is the *instantaneous rate of failure*.

It is the **(limiting) probability** that the failure event occurs in a given interval **conditional** upon survival to the beginning of that interval, divided by the width of the interval.

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{\Pr(t < T < t + \Delta t / T > t)}{\Delta t} = \frac{f(t)}{S(t)}$$

This result follows from $\Pr(A/B) = \Pr(A \cap B)/\Pr(B)$. Letting $A = t < T < t + \Delta t$ and $B = T > t$, then $A \cap B = A$, so $\Pr(t < T < t + \Delta t) = F(t + \Delta t) - F(t)$ and $\Pr(B) = S(t)$.

$$\text{Thus, } \lim_{\Delta t \rightarrow 0} \frac{\Pr(t < T < t + \Delta t / T > t)}{\Delta t} = \lim_{\Delta t \rightarrow 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \frac{1}{S(t)} = \frac{f(t)}{S(t)}.$$

Basic Concepts of Survival Analysis

The hazard rate varies from zero (meaning no risk at all) to infinity (meaning the certainty of failure at that instant).

Over time, the hazard rate can increase, decrease, remain constant – i.e. take on many different shapes.

The hazard rate measures the rate at which risk is accumulated.

For repeatable events, the hazard function is often called the **intensity function**.

“The hazard rate is at the heart of modern survival analysis...”

The human mortality pattern related to aging generates a falling hazard for a while after birth, then a long, flat plateau, and thereafter constantly rising and eventually reaching values near infinity at about 100 years.

This shape is known as the *“bathtub hazard”* by biometricians.

Chapter 2: Basic Concepts of Survival Analysis

Given any one of the four functions, $f(t)$, $F(t)$, $S(t)$, or $h(t)$, we can solve for the other three.

The **cumulative hazard function**: $H(t) = \int_0^t h(u) du$

Note that $H(t) = \int_0^t \frac{f(u)}{S(u)} du = -\int_0^t \frac{1}{S(u)} \left\{ \frac{d}{du} S(u) \right\} du = -\ln \{S(t)\}$

$$S(t) = \exp\{-H(t)\}$$

$$F(t) = 1 - \exp\{-H(t)\}$$

$$f(t) = h(t) \exp\{-H(t)\}$$

Chapter 2: Basic Concepts of Survival Analysis

From $f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}$ and $h(t) = \frac{f(t)}{S(t)}$ it follows that

$$h(t) = -\frac{d}{dt} \log S(t) \quad \text{and} \quad S(t) = \exp \left\{ -\int_0^t h(u) du \right\} \quad \text{and} \quad f(t) = h(t) \exp \left\{ -\int_0^t h(u) du \right\}$$

Thus we can move back and forth between the pdf $f(t)$ and $h(t)$.

Chapter 2: Basic Concepts of Survival Analysis

Interpretations of the Hazard Function

Some basic points about $h(t)$:

1. Being an instantaneous probability, and not an actual probability, $h(t)$ ranges from 0 to infinity.
2. Not being an observed quantity, $h(t)$ is estimated from the data.
3. $h(t)$ is observation-specific, not a population quantity.

Intuitively, $h(t)$ measures the risk of the event occurring at time t .

Its dimensions are the *number of events per unit of time*. So the units of time are important in interpreting $h(t)$.

For example, if time is measured in months and $h(t) = 0.15$ and remains constant over a month, then it would be expected that 0.15 events would occur over a one month period or 1.5 events over a 10-month period.

Chapter 2: Basic Concepts of Survival Analysis

For **repeatable events** (i.e., not like death) it makes sense to think of $h(t)$ as being the **expected number of events per unit of time**.

The reciprocal, $1/h(t)$, can be interpreted as the **expected amount of time before an event occurs**. This makes sense even for non-repeatable events.

For example, if $h(t) = 0.2$, time is measured in months, and $h(t)$ is assumed to remain constant over time, then the expected amount of time between events would be 5 months.

In most cases, $h(t)$ changes over time as situations change. For example, if the event of interest is being hit by a car, then $h(t)$ is pretty low when watching TV in the house, but increases quite a bit once you try to run across a highway.

Chapter 2: Basic Concepts of Survival Analysis

Some Simple Hazard Models

1. $h(t) = \lambda$, a **constant**, or equivalently, $\log h(t) = \mu$, $\Rightarrow S(t) = e^{-\lambda t}$

Since $f(t) = h(t) S(t) \Rightarrow f(t) = \lambda e^{-\lambda t} \Rightarrow T$ has an **exponential distribution** with parameter λ .

2. $\log h(t) = \mu + \alpha t$ (note: defining $\log h(t)$ this way ensures that $h(t) = \lambda \gamma^t$ is positive where $\lambda = e^\mu$ and $\gamma = e^\alpha$) $\Rightarrow T$ has a **Gompertz distribution**.

3. $\log h(t) = \mu + \alpha \log t \Rightarrow h(t) = \lambda t^\alpha$, $\lambda = e^\mu \Rightarrow T$ has a **Weibull distribution**.

Note that $\alpha = 0 \Rightarrow$ (2) and (3) reduce to (1).

For $\alpha > 0$ $h(t)$ is increasing and $\alpha < 0$ implies $h(t)$ is decreasing.

Chapter 2: Basic Concepts of Survival Analysis

PROC LIFEREG can estimate Weibull and exponential models.

All three models on the previous page are examples of *proportional hazards models*, which can be estimated with **PROC PHREG**.

Chapter 2: Basic Concepts of Survival Analysis

The Origin of Time

All models for survival data implicitly assume an **origin** and **scale** for the measurement of time.

Scale refers to whether time is measured in days, weeks, months, etc.

Since most models are linear in the logarithm of the $h(t)$ and t , a change in **scale** only affects the intercept.

The choice of **origin** (when to start measuring time) is much more critical.

Medical studies usually measure time of death as the length of time between the *point of diagnosis* and death.

A preferred origin would be the *initial point of infection*, but this is usually unknown.

Chapter 2: Basic Concepts of Survival Analysis

The fact that when a diagnosis is finally made could depend on factors such as race, gender, economic status, etc, could lead to seriously biased parameter estimates.

On the other hand (as an economist would say), the focus of the study might be the effectiveness of treatment, which would commence once a diagnosis was made, in which case, maybe point of diagnosis is more appropriate.

The author discusses the question of what criteria to use when choosing among several possible time origins.

Some possible choices of time origin:

Age: Demographers study age at death, in which case the birth date is a natural origin.

Chapter 2: Basic Concepts of Survival Analysis

Calendar time: When monitoring animals in the wild, it is common to choose a particular date to begin the study, say 10/1/2006, and continue the study for a fixed period of time.

Time since some other event: When studying the determinants of divorce or inmate recidivism, it makes sense to start counting time after marriage or prison release, respectively.

Time since the last occurrence of the same type of event: This only makes sense for repeatable events, such as hospitalization.

Continuous-time methods require a choice of a single time origin.

Basic Concepts of Survival Analysis

Some criteria to consider when selecting a time origin:

1. *Choose a time origin that marks the onset of continuous exposure to the risk of the event.*

See examples of divorce and recidivism above. It usually makes sense to exclude periods of time when the hazard is necessarily 0.

2. *In experimental studies, choose the time of randomization to treatment as the time origin.*

In such studies, the effect of different treatments is usually the focus. This criterion could trump the 1st criterion. For example, to study the effect of marriage counseling, the origin should be the beginning of counseling, not the date of marriage. In this case, the length of marriage could serve as a covariate.

Chapter 2: Basic Concepts of Survival Analysis

Data Structure

Most duration analysis data have a common structure. There is always a variable (called *DUR*, for example) which is the time measurement for each observation and a variable (called *STATUS*, for example) which indicates whether an observation is censored or not.

Most datasets will have additional covariates.

The next page shows a dataset used in the text. It gives survival times for 25 patients diagnosed with myelomatosis.

Chapter 2: Basic Concepts of Survival Analysis

Myelomatosis Data

Obs	id	dur	status	treat	renal
1	1	8	1	1	1
2	2	180	1	2	0
3	3	632	1	2	0
4	4	852	0	1	0
5	5	52	1	1	1
6	6	2240	0	2	0
7	7	220	1	1	0
8	8	63	1	1	1
9	9	195	1	2	0
10	10	76	1	2	0
11	11	70	1	2	0
12	12	8	1	1	0
13	13	13	1	2	1
14	14	1990	0	2	0
15	15	1976	0	1	0
16	16	18	1	2	1
17	17	700	1	2	0
18	18	1296	0	1	0
19	19	1460	0	1	0
20	20	210	1	2	0
21	21	63	1	1	1
22	22	1328	0	1	0
23	23	1296	1	2	0
24	24	365	0	1	0
25	25	23	1	2	1

Chapter 2: Basic Concepts of Survival Analysis

SAS has a variety of date-time functions that can be very useful.

For example, if the origin of an event is contained in three SAS date-time variables **ORMONTH**, **ORDAY**, and **ORYEAR** with the event time in **EVMONTH**, **EVDAY**, and **EVYEAR**, then to compute the number of days between origin and event times:

```
dur = mdy(evmonth, evday, evyear) - mdy(ormonth, orday, oryear);
```

Chapter 3: Estimating and Comparing Survival Curves with PROC LIFETEST: Introduction

Prior to 1970, the **estimation of survivor functions** was the main method of survival analysis.

Today, the major methodology is the **Cox regression model** (PROC PHREG).

PROC LIFETEST estimates survivor functions using two methods:

Kaplan-Meier method: for smaller datasets and precise event times

life-table or actuarial method – large datasets with crude time measures

Along with **computing** and **plotting** the estimated survivor function, LIFETEST provides three methods for **comparing survivor functions** for two or more groups.

LIFETEST can **test for associations** between survival time & sets of quantitative covariates.

Chapter 3: Estimating and Comparing Survival Curves with PROC LIFETEST: Kaplan-Meier Method

Kaplan-Meier method (KM) also known as the product-limit method.

The complexity of the KM estimator depends on the degree of censoring.

- 1. No censoring:** $\hat{S}(t)$ = proportion of sample observations with event time $> t$.
- 2. Single right censoring:** All censored cases are censored at the same time c and all observed event time are less than c . Then for all $t \leq c$,
 $\hat{S}(t)$ = proportion of sample observations with event time $> t$.
For $t > c$, $\hat{S}(t)$ is undefined.
- 3. Some censoring times are smaller than some event times:** $\hat{S}(t)$ is more complicated.

Chapter 3: Estimating and Comparing Survival Curves with PROC LIFETEST: Kaplan-Meier Method

In Case (3), the observed proportion of cases with event times greater than t can be biased downward because cases that are censored before t may have actually "*died*" before t .

Let $t_1 < t_2 < \dots < t_k$ be the k distinct event times.

Let n_j be the number of individuals at risk of an event at time t_j . (i.e., "*at risk*" means they have not experienced an event nor have been censored prior to time t_j .)

Let d_j be the number of individuals who *die* (i.e., experience the event) at time t_j .

The KM estimator is then defined as: $\hat{S}(t) = \prod_{j:t_j \leq t} \left[1 - \frac{d_j}{n_j} \right]$ for $t_1 \leq t \leq t_k$.

Note that $1 - \frac{d_j}{n_j}$ can be interpreted as the conditional probability that one has survived to time t_{j+1} given that one has survived to time t_j .

Chapter 3: Estimating and Comparing Survival Curves with PROC LIFETEST: Kaplan-Meier Method

Note that $\hat{S}(t)$ is constant over each interval $(t_{j-1}, t_j]$.

By definition, $S(t) = \Pr(T > t)$ where T is the random variable measuring failure time.

Thus, $S(t_1) = \Pr(T > t_1)$. Initially, $n_1 = n$ since everyone is initially at risk (assuming no late entries or failures before t_1).

$\frac{d_1}{n_1}$ is an estimate of $\Pr(T \leq t_1)$, so $\hat{S}(t_1) = 1 - \frac{d_1}{n_1} = \frac{n_1 - d_1}{n_1}$ is an estimate of $\Pr(T > t_1)$.

Given that someone has survived until t_1 , the **conditional probability** that they survive at t_2 can be estimated by $\frac{n_2 - d_2}{n_2}$.

Since $\Pr(T > t_2) = \Pr(T > t_1) \cdot \Pr(T > t_2 | T > t_1)$, thus, the unconditional probability of surviving past t_2 can be estimated by $\hat{S}(t_2) = \frac{n_1 - d_1}{n_1} \cdot \frac{n_2 - d_2}{n_2}$.

Chapter 3: Estimating and Comparing Survival Curves with PROC LIFETEST: Kaplan-Meier Method

As we move from one failure time to the next, we add another term to the product (thus the name **product limit estimator**).

The formula for the **Kaplan-Meier estimator** only involves the **failure times** and not the censored times.

The **censored observations** only affect the number of subjects at risk at any given time.

Chapter 3: Estimating and Comparing Survival Curves with PROC LIFETEST: Kaplan-Meier Method

We now use **PROC LIFETEST** to get the KM estimator for the myelomatosis dataset introduced on p. 26.

First, the dataset is sorted by *dur* and *status*.

Note that $t_1 = 8$, $t_k = 1296$, that 8 observations out of 25 are censored, and that some uncensored observations occur after some censored ones.

$\hat{S}(t)$ is undefined for $t > 1296$.

id	dur	status	treat	renal
1	8	1	1	1
12	8	1	1	0
13	13	1	2	1
16	18	1	2	1
25	23	1	2	1
5	52	1	1	1
8	63	1	1	1
21	63	1	1	1
11	70	1	2	0
10	76	1	2	0
2	180	1	2	0
9	195	1	2	0
20	210	1	2	0
7	220	1	1	0
24	365	0	1	0
3	632	1	2	0
17	700	1	2	0
4	852	0	1	0
18	1296	0	1	0
23	1296	1	2	0
22	1328	0	1	0
19	1460	0	1	0
15	1976	0	1	0
14	1990	0	2	0
6	2240	0	2	0

Chapter 3: Estimating and Comparing Survival Curves with PROC LIFETEST: Kaplan-Meier Method

```
proc lifetest data=survival.myel method=KM;  
    time dur*status(0);  
run;
```

method=KM is the default.

time dur*status(0); specifies that dur is the variable measuring event time and status is the variable indicating whether or not an observation is censored. In this case, a value of 0 indicates a censored observation.

Chapter 3: Estimating and Comparing Survival Curves with PROC LIFETEST: Kaplan-Meier Method

Product-Limit Survival Estimates						
	dur	Survival	Failure	Survival Standard Error	Number Failed	Number Left
$\hat{S}(t)$ is in the column labeled Survival .	0.00	1.0000	0	0	0	25
	8.00	.	.	.	1	24
	8.00	0.9200	0.0800	0.0543	2	23
	13.00	0.8800	0.1200	0.0650	3	22
$\hat{S}(180) = 0.56$, so the estimated probability of lasting at least 180 days is 0.56. In fact, the estimated probability of lasting anywhere between 180 and 194 days is still 0.56.	18.00	0.8400	0.1600	0.0733	4	21
	23.00	0.8000	0.2000	0.0800	5	20
	52.00	0.7600	0.2400	0.0854	6	19
	63.00	.	.	.	7	18
	63.00	0.6800	0.3200	0.0933	8	17
	70.00	0.6400	0.3600	0.0960	9	16
	76.00	0.6000	0.4000	0.0980	10	15
	180.00	0.5600	0.4400	0.0993	11	14
	195.00	0.5200	0.4800	0.0999	12	13
	210.00	0.4800	0.5200	0.0999	13	12
	220.00	0.4400	0.5600	0.0993	14	11
	365.00*	.	.	.	14	10
	632.00	0.3960	0.6040	0.0986	15	9
	700.00	0.3520	0.6480	0.0970	16	8
	852.00*	.	.	.	16	7
The last column is the size of the <i>risk set</i> , n_j , at each value of <i>dur</i> .	1296.00	0.3017	0.6983	0.0953	17	6
	1296.00*	.	.	.	17	5
	1328.00*	.	.	.	17	4
	1460.00*	.	.	.	17	3
	1976.00*	.	.	.	17	2
	1990.00*	.	.	.	17	1
	2240.00*	.	.	.	17	0

NOTE: The marked survival times are censored observations.

Chapter 3: Estimating and Comparing Survival Curves with PROC LIFETEST: Kaplan-Meier Method

For example, the 25th percentile is 63, the lowest value for *dur* for which the probability of an event occurring is at least 25%.

The **median death time**, here 210 days, is usually of greatest interest.

The estimated **mean time of death**, here 563 days, is *downward biased* when there are censored observations with time values greater than the largest observed event time.

NOTE: The mean survival time and its standard error were underestimated because the largest observation was censored and the estimation was restricted to the largest event time.

Summary Statistics for Time Variable dur

Quartile Estimates				
Percent	Point Estimate	95% Confidence Interval [Lower Upper)		
75	.	220.00	.	
50	210.00	63.00	1296.00	
25	63.00	18.00	195.00	
Mean		Standard Error		
562.76		117.32		

Summary of the Number of Censored and Uncensored Values

Total	Failed	Censored	Percent Censored
25	17	8	32.00

Chapter 3: Estimating and Comparing Survival Curves with PROC LIFETEST: Kaplan-Meier Method

To get a plot of the KM estimate (old method):

```
proc lifetest data=survival.myel plots=(s) graphics;  
  time dur*status(0);  
  symbol v=none;  
  
run;
```

