STATISTICAL HORIZONS

Survival Analysis Using Stata

Paul D. Allison, Ph.D.

Upcoming Seminar:
February 22-23, 2018, Stockholm, Sweden

Table of Contents

General Introduction

Event History Analysis = Survival Analysis = Failure-time Analysis = Reliability Analysis = Duration Analysis = Hazard Analysis = Transition Analysis

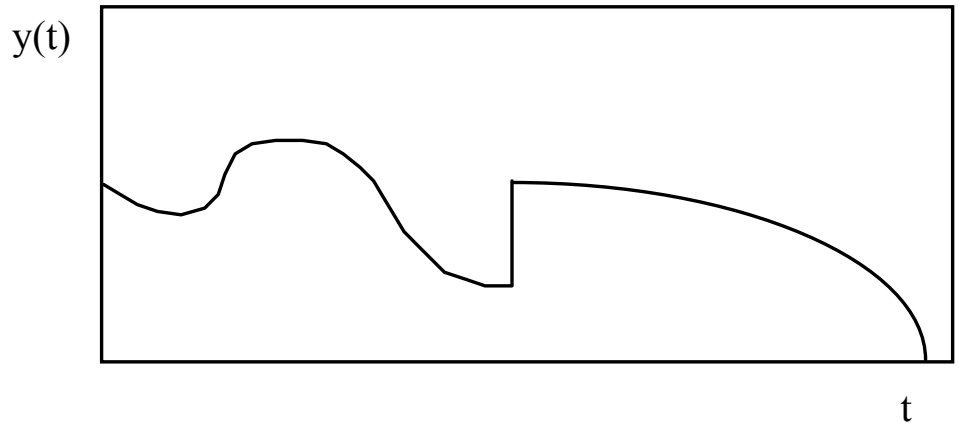
Collection of methods in which the aim is to describe how or explain why certain events do or do not occur

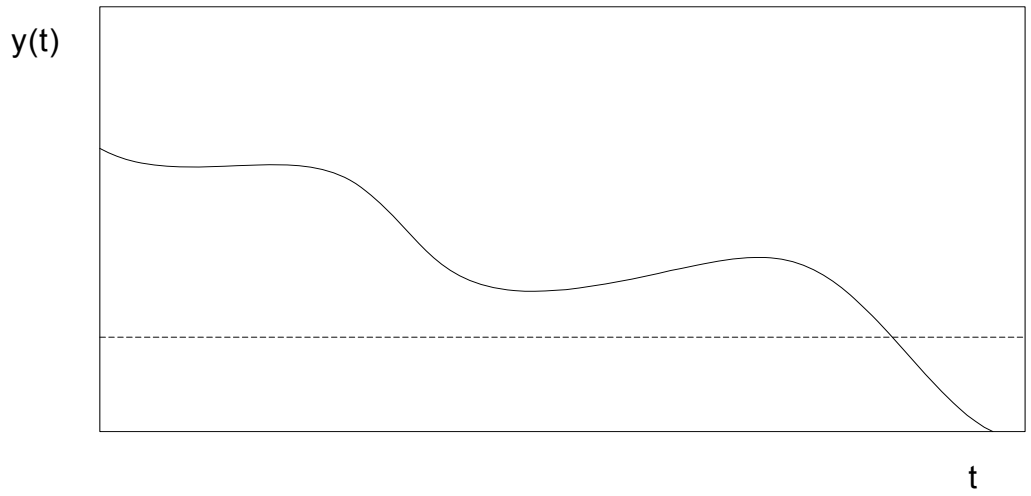
- Many different approaches
- All deal with right-censored data

Fundamentals of Event History Analysis

Event: A qualitative change that can be localized in time. Ideally, a change from one discrete state to another that occurs virtually instantaneously, e.g., death, marriage, promotion.

Can also talk about events with respect to quantitative variables so long as the change is sharp rather than gradual.





Can also define an event as occurring when some quantitative variable crosses a threshold.

Event History: A longitudinal record of when events occurred for some individual or set of individuals.

> Example: A survey which gathers retrospective information on dates of employment and unemployment.
>
> If the aim is a causal analysis, the data should also contain information on possible explanatory variables. Some of these, like sex, may be constant, while others, like income, may vary over time.

Problems with Conventional Methods

Example: 432 inmates released from Maryland state prisons, followed for one year after their release (Peter Rossi, Richard Berk, Kenneth Lenihan, *Money, Work and Crime*).

- Events: arrests.
- Explanatory variables: financial aid, education, employment status.

Method A: Dummy dependent variable (1=arrest, 0=no arrest). Linear regression on independent variables.

Problems:

- Should use logistic regression.
- Wastes information
- How to deal with employment status, which varies over time?

Method B: Dependent variable is length of time from release to first arrest.

Problems:

- What about cases with no arrests (censoring)?
- How to include time-varying explanatory variables?

Two central problems of event history data

- How to deal with censoring (how to combine information for those who did and did not experience events).
- How to incorporate explanatory variables which vary over time.

Censoring

Definitions of right, left and interval censoring:

Suppose we have a random variable T.

- We say that a particular observation of T is *right censored* if all we know about T is that it is greater than some constant c.

  example: Suppose we have a sample of women interviewed at age 30, and the event of interest is first marriage. Let T be the

age at marriage. For women still unmarried, we know only that T > 30.

- A particular observation is called *left censored* if all we know about T is that it is less than some constant c.

  example: Suppose we do a prospective study and interview women annually between the ages of 20 and 45 to determine age at first marriage. But in the initial interview, we only ask if they are currently married, not the year. For those currently married, we know only that T < 20.

- A particular observation is interval censored if all we know about T is that it is between two numbers a and b.

  example: Suppose we interview a sample of women in 2005 and 2008. At each interview, we ask their marital status. Those who are unmarried in 2005 and married in 2008 have marriage times that are interval censored.

Right censoring is far more common. In the social sciences, what is often referred to as left censoring is actually a form of right censoring. More later on this confusion.

Types of right censoring

Fixed: For each case in the sample (i = 1,...,n), there is a number c_i (determined in advance by the study design) such that if $T_i \leq c_i$ then T_i is observed, while if $T_i > c_i$ the case is censored.

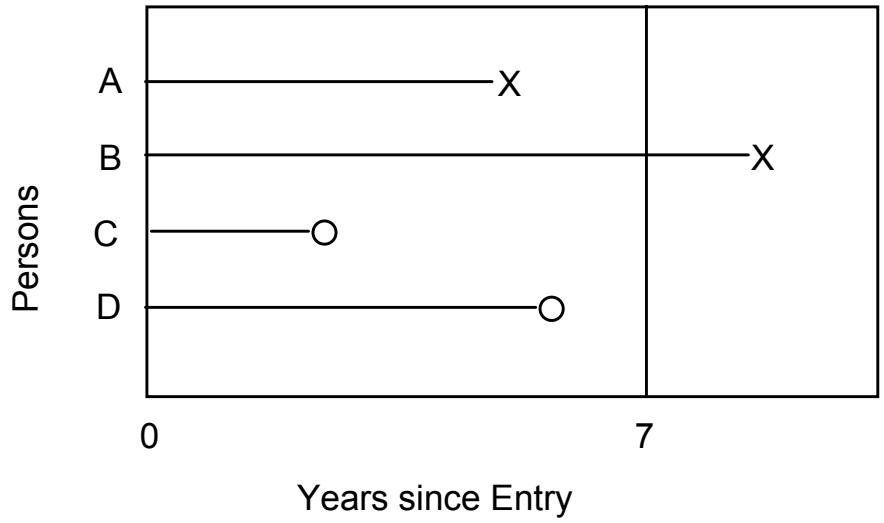
- Special case: $c_i = c$ for all i. Data are said to be "singly right censored."

- Example: Recidivism study, released inmates are followed for one year after release. T is the number of months from release to first arrest and c = 12 for all cases.

- Random Censoring: Same as Fixed, except that the c_i's are random variables rather than being determined by the study design. Occurs when individuals drop out, die, or are otherwise lost to follow-up.

Example: Sample consists of a cohort of entering sociology graduate students at the University of Pennsylvania. T is length of time from entry to receipt of Ph.D. Follow-up for seven years.

- Clearly, those who are still registered but haven't received a degree at end of seven years are censored by a Fixed mechanism. But many others will be censored at earlier times because of drop out.



Implications of censoring for analysis

- Regardless of the model being estimated, all types of censoring require special estimation procedures, usually maximum likelihood.

- Random censoring requires an additional assumption about the nature of the censoring process:

Noninformative censoring. Conditional on the explanatory variables, the fact that an individual is censored at time t does not give any information about the individual's hazard at time t. That is, individuals are not censored because they are at higher or lower risk of an event.

Possible violations:

- Graduate students withdraw because they think they don't have the ability to complete a Ph.D.

The consequences of violations are difficult to predict, but can be investigated by sensitivity analysis. The data contain no information which would help to determine whether the assumption is satisfied or not.

Describing the Distribution of Event Times.

In all approaches to event history analysis, the event time T is regarded as random or stochastic. Accordingly, we can describe it in ways that are standard for random variables.

Cumulative Distribution Function

$$F(t) = Pr(T \leq t)$$

Survivor Function

$S(t) = 1 - F(t)$, the probability of "surviving" past time t.

Density Function

$$f(t) = dF(t)/dt$$

Hazard Function

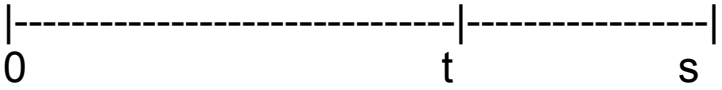
- A way of mathematically expressing the intuitive notion of the risk of event occurrence.

Definition of the hazard function (rate, hazard, intensity, hazard rate function, force of mortality)

Let T be the random variable denoting the time of event occurrence.

Let $P(t, s) = Pr(t < T < s \mid T \geq t)$,

i.e., the probability that an event occurs between t and s given that an event has not already occurred.



Then,

$$h(t) = \lim_{s \to t} \frac{P(t, s)}{s - t}$$

Other symbols: r(t), $\lambda(t)$

Interpretation of the hazard:

- Like a probability, a hazard is never directly observed
- "Instantaneous probability of event occurrence"
  Not really appropriate because h(t) can be greater than one.
  No upper bound.

- If h(t) is a constant c, then c is the expected number of events in an interval that is one time-unit long. Thus the scale of measurement is events / time.

  e.g. h(t) = .78 for all t. Then we expect .78 events per unit time.

  Example: We observe 10,000 individuals, all of whom are exactly the same, for a period of one year, during which the hazard is constant. 200 of them die.

  Let U be the total amount of time (in years) that all individuals are observed. For all those who don't die, the contribution to U is 9,800. For those who do die, the contribution to U is the length of time from the beginning of the year to the time of death.

  Then an optimal estimate of the hazard is 200/U.

- Alternatively, if h(t) is constant over t, then E(T) = 1/h(t).

  e.g. h(t) = .78 for all t and time is measured in years, then 1/.78 = 1.28 years is expected length of time until an event occurs.

Note: We let the hazard be a function of t so that the instantaneous risk can vary with time.

Types of hazard functions

- Initially assume no explanatory variables: Model is for a single individual or a set of homogeneous individuals.
- Each type of hazard function defines a family of distributions for T, i.e., hazard functions are a way of characterizing a distribution.
- There are equivalences between hazard functions and other ways of describing distributions, thus:

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}$$

$$f(t) = h(t) \exp\left(- \int_0^t h(u) du \right)$$

$$F(t) = 1 - \exp\left(- \int_0^t h(u) du \right)$$

where f(t) is the probability density function and F(t) is the cumulative distribution function.

(a) constant hazard (exponential model)

- Implies that T has an exponential distribution:
  Suppose h(t) = λ for all t, then density for T is

  $f(t) = \lambda \exp(-\lambda t)$

- Often useful as a baseline model, simplifies calculations

(b) Gompertz model

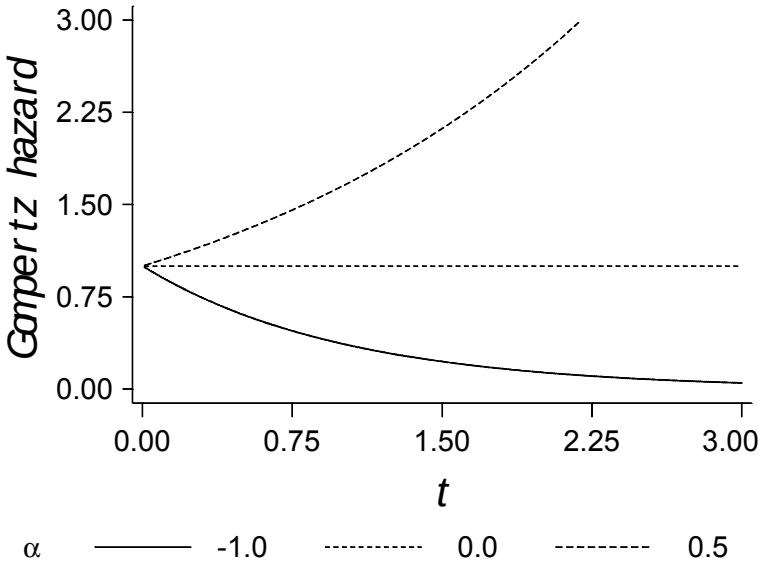
$\log h(t) = \mu + \alpha t$

where α can be positive or negative.

- Why log? Because $h(t) \geq 0$.
- Note: All logarithms to the base e = 2.71828....
- Equivalently

  $h(t) = \lambda_0 \exp\{\alpha t\}$ where $\lambda_0 = e^{\mu}$.

- Implies that T has a Gompertz distribution



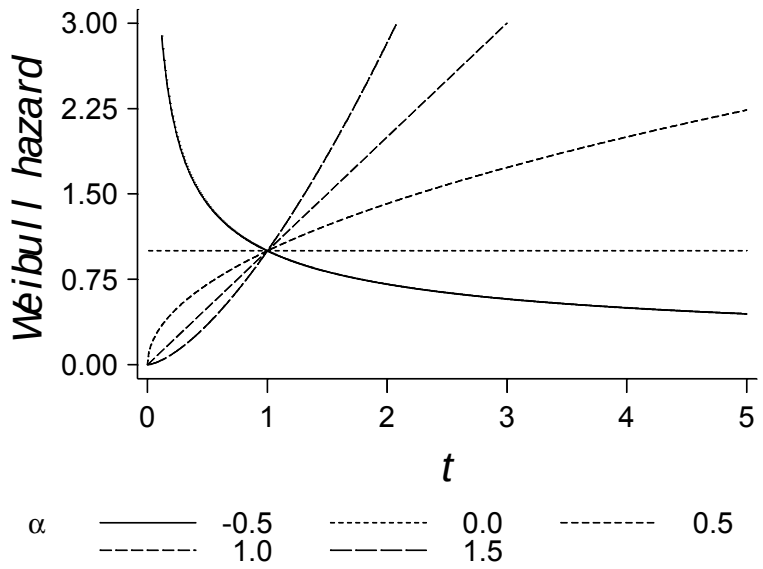
(c) Weibull model

$$\log h(t) = \mu + \alpha \log t \text{ where } \alpha > -1$$

Equivalently,

$$h(t) = \lambda_0 t^{\alpha}$$

- Implies that T has a Weibull distribution



Inclusion of explanatory variables:

$$\log h(t) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \qquad \text{(exponential)}$$

$$\log h(t) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \alpha t \qquad \text{(Gompertz)}$$

$$\log h(t) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \alpha \log t \qquad \text{(Weibull)}$$

- All three models are members of a general class of models known as proportional hazards models.

- Weibull (and exponential) is both a proportional hazards model and an accelerated failure-time model. It is unique in this respect.

Estimation: Maximum likelihood (using **streg**).

Nonparametric Estimation of Survivor Functions

Let S(t) = Pr(T > t) where T is the time of the event, i.e., the probability that an individual "survives" to time t. Recall that S(t) = 1 - F(t) where F(t) is the c.d.f.

$0 \leq S(t) \leq 1$, a non-increasing function of t.

Every distribution for T has a corresponding survival curve.

When there is no censoring, it's easy to estimate the survivor function. For any t, simply estimate S(t) by the proportion of cases that have survived past that point in time.

If all event times are less than all censored times, the survivor function can again be estimated by the proportion surviving, for all times up to the lowest censored time. After that the estimate is undefined (clearly it must be between the last value and 0).

If some censored times are less than some event times, more complex methods are needed. The standard methods are the life table method and the Kaplan Meier method.

Kaplan-Meier Method

Example: 65 myeloma patients

DUR: time (in months) from diagnosis to death or censoring.
NIT: log of blood urea nitrogen at diagnosis
HEMO: hemoglobin at diagnosis
AGE: age at diagnosis
SEX: 0=male, 1=female
CALC: blood calcium at diagnosis
DEAD: 1=uncensored, 0 = censored.

17 cases are censored.

OBS	DUR	NIT	HEMO	AGE	SEX	CALC	DEAD
1	1	2.218	9.4	67	0	10	1
2	1	1.940	12.0	38	0	18	1
3	2	1.519	9.8	81	0	15	1
4	2	1.748	11.3	75	0	12	1

5	2	1.301	5.1	57	0	9	1
6	3	1.544	6.7	46	1	10	1
7	5	2.236	10.1	50	1	9	1
8	5	1.681	6.5	74	0	9	1
9	6	1.362	9.0	77	0	8	1
10	6	2.114	10.2	70	1	8	1
11	6	1.114	9.7	60	0	10	1
12	6	1.415	10.4	67	1	8	1
13	7	1.978	9.5	48	0	10	1
14	7	1.041	5.1	61	1	10	1
15	7	1.176	11.4	53	1	13	1
16	9	1.724	8.2	55	0	12	1
17	11	1.114	14.0	61	0	10	1
18	11	1.230	12.0	43	0	9	1
19	11	1.301	13.2	65	0	10	1
20	11	1.508	7.5	70	0	12	1
21	11	1.079	9.6	51	1	9	1
22	13	0.778	5.5	60	1	10	1
23	14	1.398	14.6	66	0	10	1
24	15	1.602	10.6	70	0	11	1
25	16	1.342	9.0	48	0	10	1
26	16	1.322	8.8	62	1	10	1
27	17	1.230	10.0	53	0	9	1
28	17	1.591	11.2	68	0	10	1
29	18	1.447	7.5	65	1	8	1
30	19	1.079	14.4	51	0	15	1
31	19	1.255	7.5	60	1	9	1
32	24	1.301	14.6	56	1	9	1
33	25	1.000	12.4	67	0	10	1
34	26	1.230	11.2	49	1	11	1
35	32	1.322	10.6	46	0	9	1
36	35	1.114	7.0	48	0	10	1
37	37	1.602	11.0	63	0	9	1
38	41	1.000	10.2	69	0	10	1
39	42	1.146	5.0	70	1	9	1
40	51	1.568	7.7	74	0	13	1
41	52	1.000	10.1	60	1	10	1
42	54	1.255	9.0	49	0	10	1
43	58	1.204	12.1	42	1	10	1
44	66	1.447	6.6	59	0	9	1
45	67	1.322	12.8	52	0	10	1
46	88	1.176	10.6	47	1	9	1
47	89	1.322	14.0	63	0	9	1
48	92	1.431	11.0	58	1	11	1
49	4	1.945	10.2	59	0	10	0
50	4	1.924	10.0	49	1	13	0

51	7	1.114	12.4	48	1	10	0
52	7	1.532	10.2	81	0	11	0
53	8	1.079	9.9	57	1	8	0
54	11	1.613	14.0	60	0	9	0
55	12	1.146	11.6	46	1	7	0
56	12	1.398	8.8	66	1	9	0
57	13	1.663	4.9	71	1	9	0
58	16	1.146	13.0	55	0	9	0
59	19	1.322	13.0	59	1	10	0
60	19	1.322	10.8	69	1	10	0
61	28	1.230	7.3	82	1	9	0
62	41	1.756	12.8	72	0	9	0
63	53	1.114	12.0	66	0	11	0
64	57	1.255	12.5	66	0	11	0
65	77	1.079	14.0	60	0	12	0

Since time of death is measured in months, let q_j be the conditional probability of dying in month j given that you were still alive at the beginning of the month. That can be estimated by

$$\hat{q}_j = \frac{d_j}{n_j}$$

where d_j is the number dying in month j and n_j is the number still alive at the beginning of the month. (If you're censored in month j, you're still in the denominator).

If we want the probability of surviving past, say, the 50th month, an appropriate estimate is

$$(1-\hat{q}_1)(1-\hat{q}_2)(1-\hat{q}_3)....(1-\hat{q}_{50})$$

For 27 of these 50 months, however, no deaths occurred. So the estimate for q on those months is 0 and, of course, 1-q = 1, which has no effect on the estimate. So we really only need to calculate terms for the months on which events occurred.

The general form of the KM estimator is

$$\hat{S}(t) = \prod_{j|t_j \le t} \left(1 - \frac{d_j}{n_j} \right)$$

To implement the KM method in Stata, first declare your data to be survival time data with the **stset** command:

```
use c:\data\myel.dta, clear
stset dur, failure(dead==1)
```

```
     failure event:  dead == 1
obs. time interval:  (0, dur]
 exit on or before:  failure

--------------------------------------------------------------------------
       65  total obs.
        0  exclusions
--------------------------------------------------------------------------
       65  obs. remaining, representing
       48  failures in single record/single failure data
     1561  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =        92
```

Then, to get the KM estimator, use

```
sts list
```

```
        failure _d:  dead == 1
  analysis time _t:  dur
```

Time	Beg. Total	Fail	Net Lost	Survivor Function	Std. Error	[95% Conf.	Int.]
1	65	2	0	0.9692	0.0214	0.8825	0.9922
2	63	3	0	0.9231	0.0331	0.8250	0.9672
3	60	1	0	0.9077	0.0359	0.8060	0.9574
4	59	0	2	0.9077	0.0359	0.8060	0.9574
5	57	2	0	0.8758	0.0411	0.7670	0.9359
6	55	4	0	0.8121	0.0489	0.6928	0.8887
7	51	3	2	0.7644	0.0533	0.6398	0.8507
8	46	0	1	0.7644	0.0533	0.6398	0.8507
9	45	1	0	0.7474	0.0547	0.6209	0.8370
11	44	5	1	0.6625	0.0603	0.5300	0.7656
12	38	0	2	0.6625	0.0603	0.5300	0.7656
13	36	1	1	0.6441	0.0613	0.5105	0.7499
14	34	1	0	0.6251	0.0624	0.4905	0.7336
15	33	1	0	0.6062	0.0633	0.4707	0.7171
16	32	2	1	0.5683	0.0648	0.4321	0.6834
17	29	2	0	0.5291	0.0660	0.3928	0.6481
18	27	1	0	0.5095	0.0664	0.3736	0.6301
19	26	2	2	0.4703	0.0668	0.3359	0.5936
24	22	1	0	0.4489	0.0671	0.3152	0.5737
25	21	1	0	0.4275	0.0672	0.2949	0.5536
26	20	1	0	0.4062	0.0672	0.2750	0.5333
28	19	0	1	0.4062	0.0672	0.2750	0.5333
32	18	1	0	0.3836	0.0671	0.2540	0.5118
35	17	1	0	0.3610	0.0669	0.2334	0.4900
37	16	1	0	0.3385	0.0664	0.2134	0.4678
41	15	1	1	0.3159	0.0657	0.1938	0.4453
42	13	1	0	0.2916	0.0650	0.1727	0.4212
51	12	1	0	0.2673	0.0639	0.1523	0.3966
52	11	1	0	0.2430	0.0626	0.1325	0.3715
53	10	0	1	0.2430	0.0626	0.1325	0.3715
54	9	1	0	0.2160	0.0612	0.1107	0.3441
57	8	0	1	0.2160	0.0612	0.1107	0.3441
58	7	1	0	0.1851	0.0597	0.0860	0.3137
66	6	1	0	0.1543	0.0572	0.0635	0.2817
67	5	1	0	0.1234	0.0534	0.0434	0.2479
77	4	0	1	0.1234	0.0534	0.0434	0.2479
88	3	1	0	0.0823	0.0490	0.0186	0.2089
89	2	1	0	0.0411	0.0380	0.0036	0.1639
92	1	1	0	0.0000	.	.	.

To get the median time to event, use

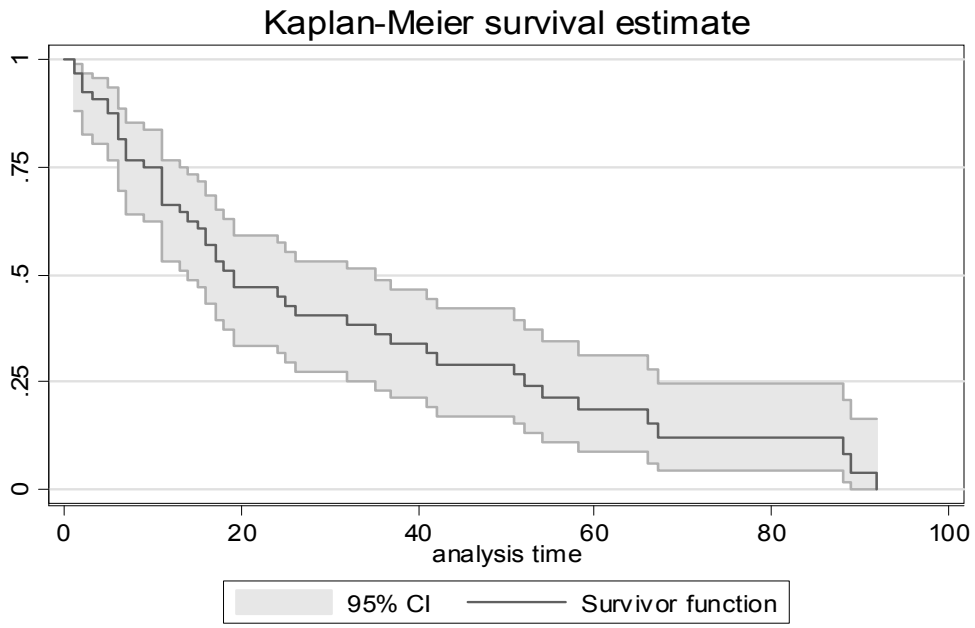
```
stci
```

```
      failure _d:  dead
analysis time _t:  dur

           |    no. of
           |  subjects         50%     Std. Err.     [95% Conf. Interval]
-----------+-------------------------------------------------------------
       total |        65          19      4.477061           14          35
```

To get a graph of the survivor function, use

```
sts graph, ci
```



For large data sets with many event times, you can reduce the output greatly by using the “at” option. This allows you to specify time values at which you want the survivor function *reported* (doesn’t change the actual estimates).