Lecture 6: Survival Analysis

Dankmar Böhning

Southampton Statistical Sciences Research Institute
University of Southampton

S^{3}RI, 2 - 4 March 2015

A couple of questions and...

- What makes survival data so special that their analysis needs a special treatment, even as long as a one-term course?
- Why isn't it simply covered as a sub-topic in, let's say, regression analysis?

...a clarification

- Survival data subsume more than only times from birth to death for some individuals.
- Analysis of duration data, that is the time from a well-defined starting point until the event of interest occurs.

Examples

- how long patients *survived* after diagnosis or treatment
- the length of unemployment spells
- how long a marriage lasts
- how long PhD students need to finish writing their theses
- and more...

Features

- Survival data result from a dynamic process and we want to capture these dynamics in the analysis properly.
- The observation scheme for duration data can be rather complex, leading to data that are somehow *cut*.

The basic functions

In the following we will assume that time is running continuously, and we therefore will describe duration by a continuous random variable, denoted by T.

- $T \geq 0$
- $f(t) \Rightarrow$ density function
- $F(t) \Rightarrow$ cumulative density function (cdf)
- $S(t) \Rightarrow$ survival function

Recall that...

- The density function $f(t)$ describes how the total probability of 1 is distributed over the domain of T.
- The function $f(t)$ itself is not a probability and can take values bigger than 1. But still one can derive basic properties from looking at the density.
- For regions where the density has large values the area under the curve over an interval of given length will be larger as compared to an interval of same length where the density is lower.
- Regions over which the density is high are regions where we expect to observe more data points than in regions with low densities.

Recall that...

- The cdf $F(t)$ is defined as $F(t) := P(T \leq t)$ which can be computed from the density as

$$F(t) = \int_0^t f(s)ds$$

.

- A cdf is an increasing function, even strictly increasing if the density $f(t) > 0$ everywhere.
- $F(0) = 0$ and $\lim_{t\to\infty} F(t) = 1$.
- There is a one-to-one link between $f(t)$ and $F(t)$ as $F'(t) = f(t)$. Knowing one of the functions means, at least in principle, knowing the other (you may have to take the derivative or perhaps solve an *ugly* integral).

Recall that...

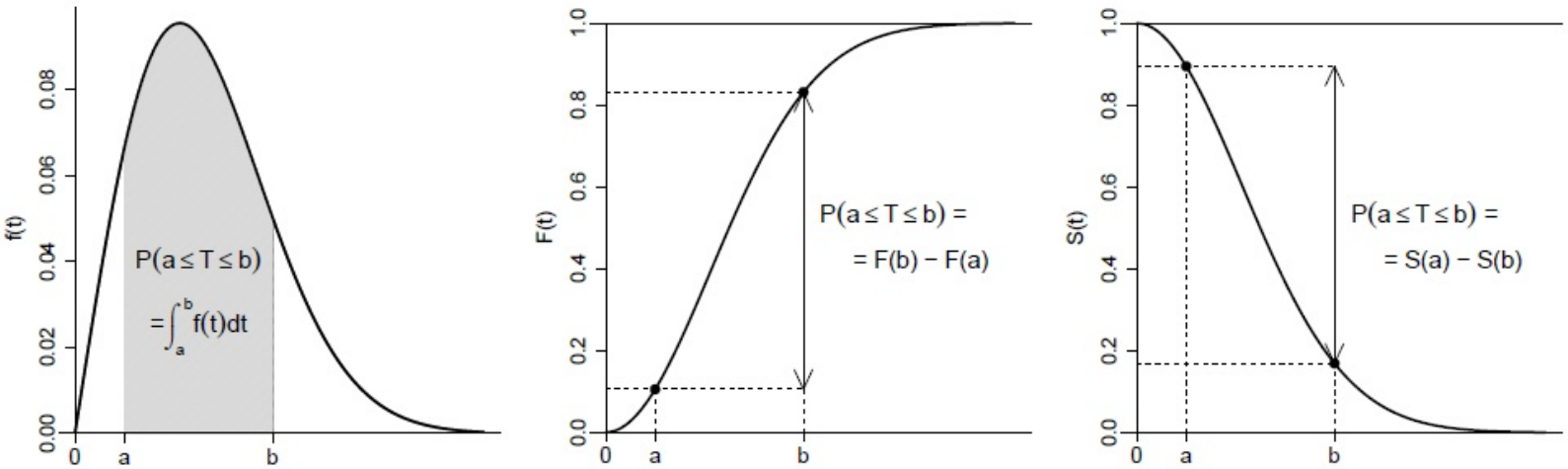
- Instead of looking at the cdf, which gives the probability of surviving at most t time units, one prefers to look at survival beyond a given point in time. This is described by the survival function $S(t)$:

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t)$$

- Consequently, $S(t)$ starts at 1 for $t = 0$ and then declines to 0 for $t \to \infty$.
- It should be obvious that knowing any one of $f(t)$, $F(t)$ and $S(t)$ allows to derive the other two functions.

To summarize

$$\Pr(a \leq T \leq b)$$



All the three functions introduced so far allowed to describe, in one way or another, how the survival times are distributed over the potential range.

The dynamic process

- Density, cdf and survival function look at the marginal distribution
- Conditioning on the survival experience so far, we have

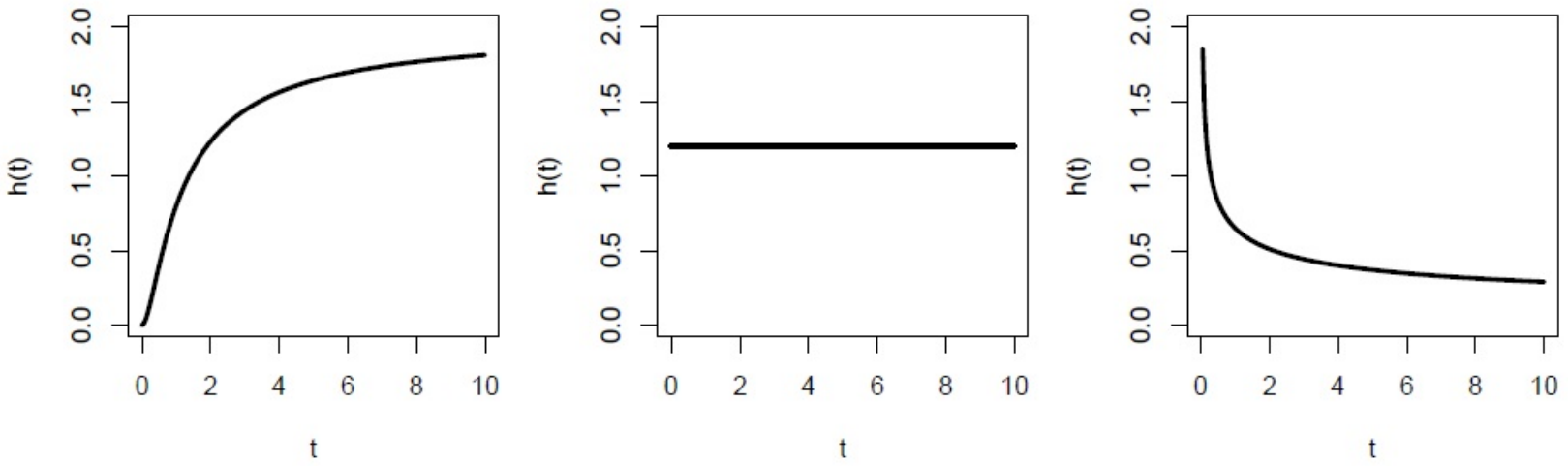
$$\Pr(t < T \leq t + \Delta t \mid T > t)$$

- Defining the Hazard Rate

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t < T \leq t + \Delta t \mid T > t)}{\Delta t}$$

The hazard in more details

The basic information in the hazard is, first of all, its qualitative behavior.



Some useful identities

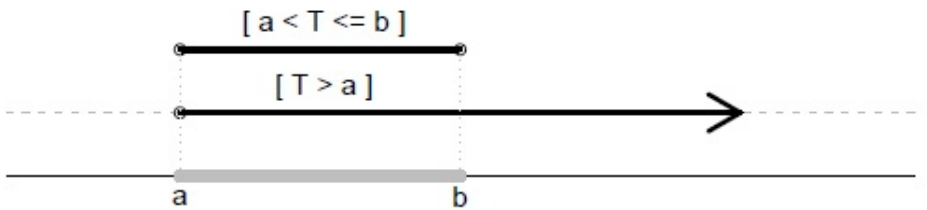
- $h(t) = \frac{f(t)}{S(t)} \Rightarrow f(t) = h(t)S(t)$
- $h(t) = [-\log S(t)]'$
- $S(t) = exp\left\{-\int_0^t h(s)ds\right\}$
- Define the cumulative hazard $H(t)$

$$H(t) = \int_0^t h(s)ds \Rightarrow S(t) = exp\{-H(t)\} or \log S(t) = -H(t)$$

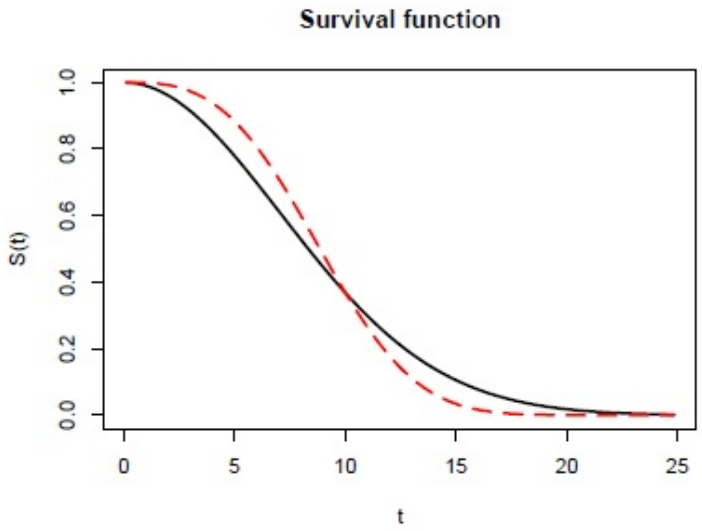
By using the definition of conditional probabilities

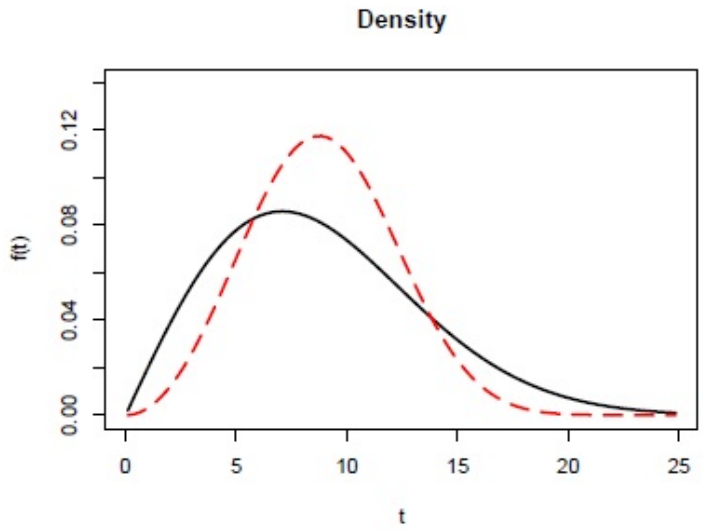
$$\begin{aligned}\Pr(t < T \leq t + \Delta t \mid T > t) &= \frac{\Pr([t < T \leq t + \Delta t] \cap [T > t])}{\Pr(T > t)} \\ &= \frac{\Pr(t < T \leq t + \Delta t \mid T > t)}{\Pr(T > t)}\end{aligned}$$

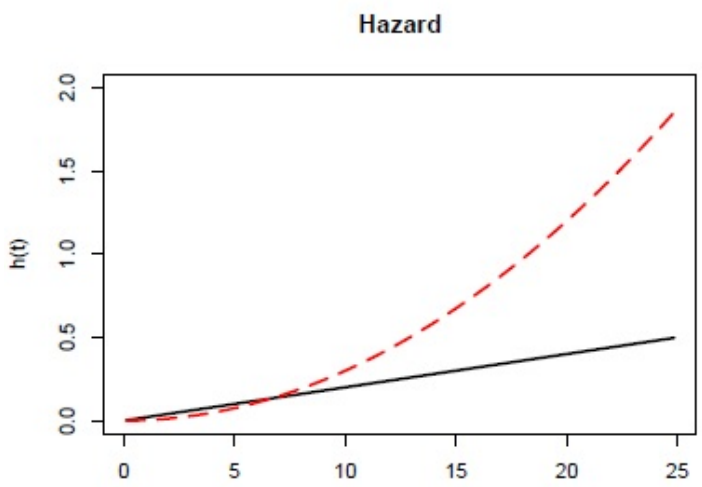
It may be helpful to sketch this relation graphically

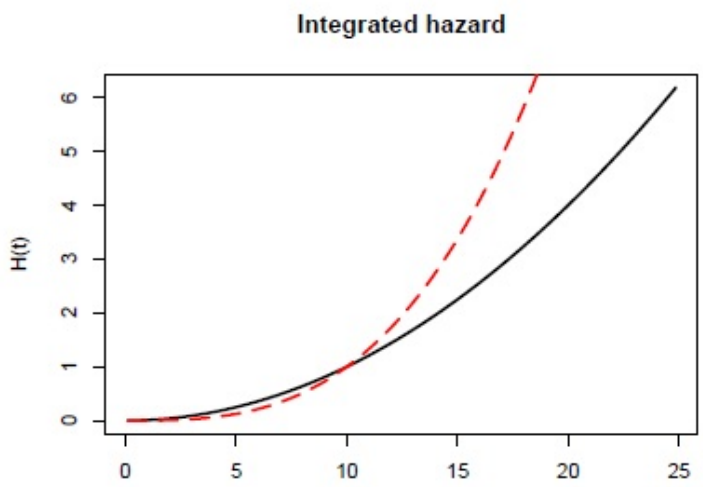


An example


Survival function
S(t)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
t


Density
f(t)
0.00
0.04
0.08
0.12
0
5
10
15
20
25
t


Hazard
h(t)
0.0
0.5
1.0
1.5
2.0
0
5
10
15
20
25


Integrated hazard
H(t)
0
1
2
3
4
5
6
0
5
10
15
20
25

Survival Analysis: Non-Parametric Estimation

Dankmar Böhning

Southampton Statistical Sciences Research Institute
University of Southampton

S^3RI, 2 - 4 March 2015

Few remarks before starting

- Each subject has a beginning and an end anywhere along the time line of the complete study.
- In many clinical trials, subjects may enter or begin the study and reach end-point at vastly differing points.
- Each subject is characterized by
  1. Survival time
  2. Status at the end of the survival time (event occurrence or censored)
  3. The study group they are in.

Censoring

- The total survival time for that subject cannot be accurately determined.
  - A subject drops out, is lost to follow-up, or required data are not available
  - The study ends before the subject had the event of interest occur, i.e., they survived at least until the end of the study,
- There is no knowledge of what happened thereafter.

Censoring

- Right censoring: the period of observation expires, or an individual is removed from the study, before the event occurs.
- Left censoring: the initial time at risk is unknown.
- Interval censoring: both right and left censored

Estimation

- Random variable T with cdf $F(t)$
- $S(t) = 1 - F(t)$
- With no censored observations:

$$\hat{S}(t) = 1 - \hat{F}(t)$$

- To estimate $F(t)$ at each time t:
  - data $t_1, \ldots, t_n$
  - parameter of interest $\theta = F(t) = \Pr(T \leq t)$
  - $\hat{\theta} = \frac{\#obs. \leq t}{n} = \frac{\sum_{i=1}^{n} \mathcal{I}_{(0,t_i)}(t)}{n}$

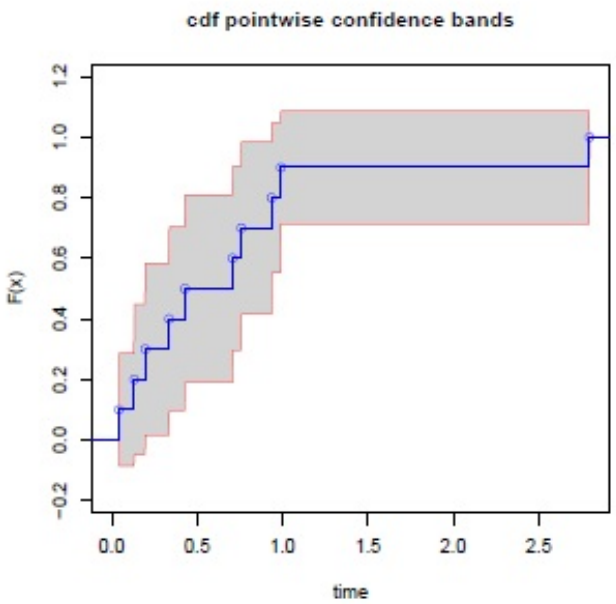
Confidence intervals

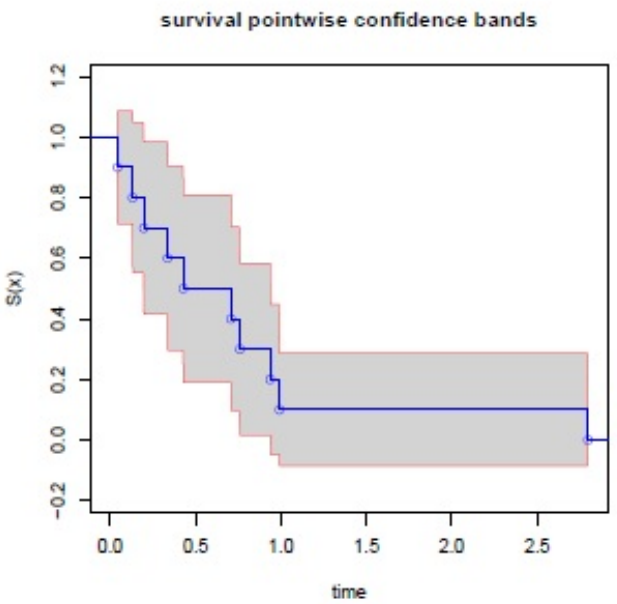
- Confidence interval for $F(t)$:

$$\hat{\theta} \mp z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$

- Confidence interval for $S(t)$:

$$1-\hat{\theta} \mp z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$


cdf pointwise confidence bands
F(x)
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0.0
0.5
1.0
1.5
2.0
2.5
time


survival pointwise confidence bands
S(x)
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0.0
0.5
1.0
1.5
2.0
2.5
time

Estimation

- To estimate the proportions θ_i
  - $n_i = \#$ of individuals at risk at the beginning of the i-th interval
  - $d_i = \#$ of individuals experiencing the event

$$\hat{\theta}_i = \frac{n_i - d_i}{n_i}$$

- Kaplan Meier estimator

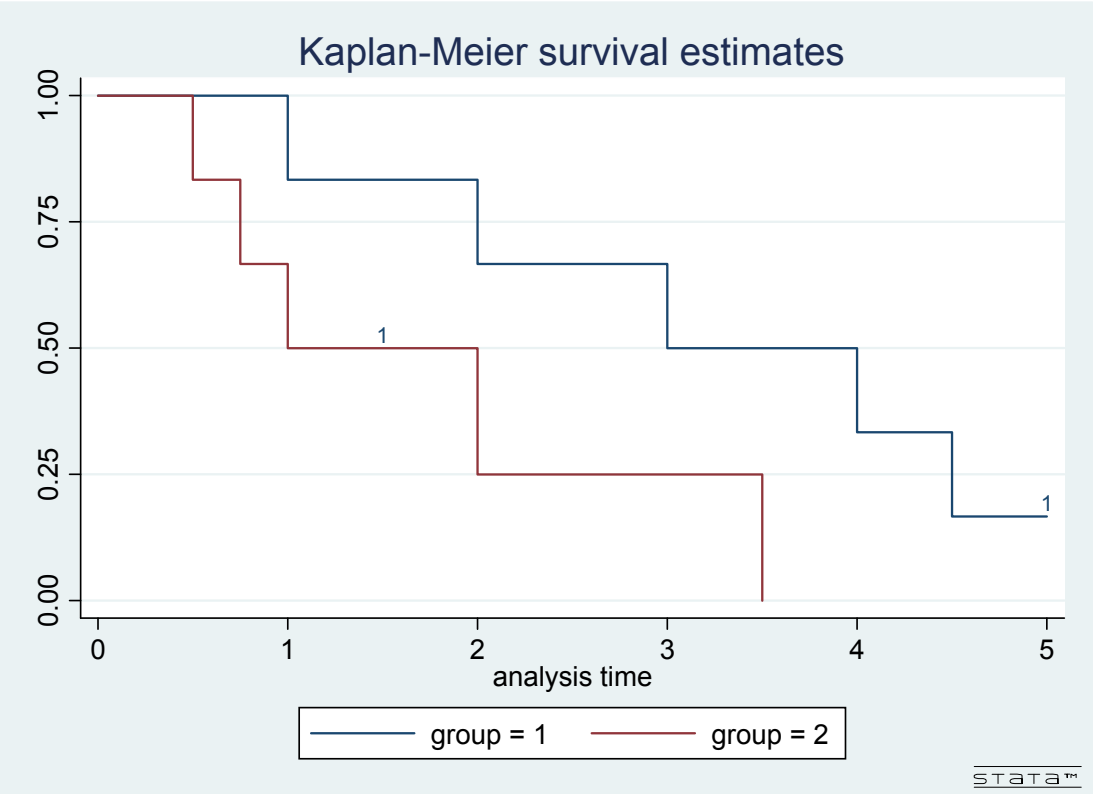
$$\hat{S}(t) = \prod_{i:t_i \leq t} \frac{n_i - d_i}{n_i}$$

- It reduces to $1 - \hat{F}(t)$ with no censored observations

Example

Subject	Group	Survival time in the interval	# *surviving* at risk	Event	# *surviving* after event	Cumulative survival rate
1	1	1	6	1	5	$1\times\frac{5}{6}$
2	1	2	5	1	4	$1\times\frac{5}{6}\times\frac{4}{5}$
3	1	3	4	1	3	$1\times\frac{5}{6}\times\frac{4}{5}\times\frac{3}{4}$
4	1	4	3	1	2	$1\times\frac{5}{6}\times\frac{4}{5}\times\frac{3}{4}\times\frac{2}{3}$
5	1	4.5	2	1	1	$1\times\frac{5}{6}\times\frac{4}{5}\times\frac{3}{4}\times\frac{2}{3}\times\frac{1}{2}$
6	1	5		0		
7	2	0.5	6	1	5	$1\times\frac{5}{6}$
8	2	0.75	5	1	4	$1\times\frac{5}{6}\times\frac{4}{5}$
9	2	1	4	1	3	$1\times\frac{5}{6}\times\frac{4}{5}\times\frac{3}{4}$
10	2	1.5		0		
11	2	2	2	1	1	$1\times\frac{5}{6}\times\frac{4}{5}\times\frac{3}{4}\times\frac{1}{2}$
12	2	3.5	1	1	0	$1\times\frac{5}{6}\times\frac{4}{5}\times\frac{3}{4}\times\frac{1}{2}$

Example


Kaplan-Meier survival estimates
1.00
0.75
0.50
0.25
0.00
1
1
0
1
2
3
4
5
analysis time
group = 1
group = 2
Stata™

Understanding KM analysis

- The lengths of the horizontal lines represent the survival duration for that interval.
- The interval is terminated by the occurrence of the event of interest.
- The vertical distances between horizontal lines illustrate the change in the cumulative probability.
- The KM curve is a step-wise estimator, not a smooth function.
- What about estimate of point survival?
- Which is the effect of censoring?

Comparison of KM estimates

- It is simple to visualize the difference between two survival curves.
- The difference must be quantified in order to assess statistical significance.
- Methods
  - log-rank test $\Rightarrow$ Most sensitive to consistent difference
  - Wilcoxon test $\Rightarrow$ Most sensitive to early differences
  - hazard ratio $\Rightarrow$ gives relative event rate in the groups

Log-Rank test: Example

Time	Group 1 Event	Group 2 Event	Group 1 At Risk	Group 2 At Risk	Group 1 Expected	Group 2 Expected
0.5	0	1	6	6	0.50	0.50
0.75	0	1	6	5	0.55	0.45
1	1	1	6	4	1.20	0.80
2	1	1	5	2	1.43	0.57
3	1	0	4	1	0.80	0.20
3.5	0	1	3	1	0.75	0.25
4	1	0	3	0	1.00	0.00
4.5	1	0	2	0	1.00	0.00

The logrank test statistic is constructed by computing the observed and expected number of events in one of the groups at each observed event time and then adding these to obtain an overall summary across all time points where there is an event.

$$\chi^2 = 3.07; p - value = 0.0798$$

What to avoid

- Compare mean survival $\Rightarrow$ Censoring makes this meaningless
- Overinterpret the tail of a survival curve $\Rightarrow$ There are generally few subjects in tails
- Compare proportions surviving at a fixed time $\Rightarrow$ Fine for description, not for hypothesis testing

Cox Proportional Hazards Regression for Survival Data

Dankmar Böhning

Southampton Statistical Sciences Research Institute
University of Southampton

S^{3}RI, 2 -4 March 2015

Survival distributions

- Survival analysis focuses on the distribution of survival times.
- Although there are well known methods for estimating unconditional survival distributions, most interesting survival modeling examines the relationship between survival and one or more predictors.
- In principle, every distribution on $\mathbb{R}^+$ can serve to characterize survival data.
  - Constant hazard
  - Gompertz distribution
  - Weibull distribution

Survival distributions

Modeling of survival data usually employs the hazard function

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t < T \leq t + \Delta t \mid T > t)}{\Delta t}$$

- Constant hazard: $h(t) = \lambda \Rightarrow S(t) = e^{-\lambda t}$
- Gompertz: $h(t) = ae^{bt}, a > 0, b > 0 \Rightarrow S(t) = e^{\frac{a}{b}[1-e^{bt}]}$
- Weibull: $h(t) = \lambda a t^{a-1} \Rightarrow S(t) = e^{-\lambda t^a}$

Regression-like model

A parametric model based on the exponential distribution may be written as

$$\log h_i(t) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

log-baseline hazard

The constant β_0 represents a kind of log-baseline hazard

The Cox model

The Cox model leaves the baseline hazard function $\beta_0(t) = \log h_0(t)$ unspecified

$$\log h_i(t) = \beta_0(t) + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

The model is semiparametric, because while the baseline hazard can take any form, the covariates enter the model linearly.

- The baseline hazard does not depend on covariates, but only on time
- The covariates are time-constant
- Proportional hazard assumption follows

The hazard ratio

For two observations i and j, the hazard ratio

$$\begin{aligned}\frac{h_i(t)}{h_j(t)} &= \frac{h_0(t)\exp(\beta_1 x_{i1} + \cdots + \beta_p x_{ip})}{h_0(t)\exp(\beta_1 x_{j1} + \cdots + \beta_p x_{jp})} \\ &= \frac{\exp(\beta_1 x_{i1} + \cdots + \beta_p x_{ip})}{\exp(\beta_1 x_{j1} + \cdots + \beta_p x_{jp})} \\ &= \exp\left(\sum_{l=1}^{p} \beta_l (x_{il} - x_{jl})\right)\end{aligned}$$

is independent of time t. Consequently, the Cox model is a proportional hazards model.

The hazard ratio: an example

- Only one covariate: Treatment
  - $x_i = 1 \Rightarrow$ Placebo
  - $x_j = 0 \Rightarrow$ Treatment
- Hazard ratio is then $\exp(\beta_1)$
- We expect that hazard is larger in the placebo group, i.e. the hazard ratio is expected grater than 1.

Time-constant covariates

- Not changing over time (e.g. gender)
- Values are set at time $t = 0$
- Variables unlikely to change are often considered time-constant
- Other variables are sometimes treated as time independent
- Time-dependent covariates are allowed, but PH assumptions is not satisfied (an extended Cox model is needed)

Advantages

- Robustness
- Because of the model form, the estimated hazards are always non-negative
- We can estimate fixed effects and compute the hazard ratio even though the baseline hazard is left unspecified

Checking proportional hazards

- Test and graphical diagnostic for PH may be based on **scaled Schoenfeld residuals**
- Influential observations
- Nonlinearity

Survival Analysis in STATA

Dankmar Böhning

Southampton Statistical Sciences Research Institute
University of Southampton

S^{3}RI, 2 - 4 March 2015

Aim

Illustrate how to use Stata to

- prepare survival data for analysis
- estimate hazard and survival functions

Data manipulation

A *manipulation* of the data is needed to facilitate summary and analysis.

```
help st
```

```
The st commands are

    stset                Declare data to be survival-time data
    stdescribe           Describe survival-time data
    stsum                Summarize survival-time data
    stvary               Report whether variables vary over time
    stfill               Fill in by carrying forward values of
                           covariates
    stgen                Generate variables reflecting entire histories
    stsplit              Split time-span records
    stjoin               Join time-span records
    stbase               Form baseline dataset
    sts                  Generate, graph, list, and test the survivor
                           and cumulative hazard functions
    stir                 Report incidence-rate comparison
    stci                 Confidence intervals for means and percentiles
                           of survival time
    strate               Tabulate failure rate
    stptime              Calculate person-time
    stmh                 Calculate rate ratios with the Mantel-Haenszel
                           method
    stmc                 Calculate rate ratios with the Mantel-Cox
                           method
    stcox                Fit Cox proportional hazards model
    estat concordance    Calculate Harrell's C
    estat phtest         Test Cox proportional-hazards assumption
    stphplot             Graphically assess the Cox proportional-hazards
                           assumption
    stcoxkm              Graphically assess the Cox proportional-hazards
                           assumption
    streg                Fit parametric survival models
    stcurve              Plot survivor, hazard, or cumulative hazard
                           function
    stpower              Sample-size, power, and effect-size
                           determination for survival studies
    stpower cox          Sample size, power, and effect size for the Cox
                           proportional hazards model
    stpower exponential  Sample size and power for the exponential test
    stpower logrank      Sample size, power, and effect size for the
                           log-rank test
    sttocc               Convert survival-time data to case-control data
    sttoct               Convert survival-time data to count-time data
    st_*                 Survival analysis subroutines for programmers
```

Assumptions

- Continuous time survival data
- Single failure data, i.e. one record per unit
- No complications such as truncation and/or missing values
- Data do not need to be weighted

Data structure

Data have a very simple structure

- One row per unit (e.g. subject)
- The survival time and the censoring status must be included as variables (1= failure, 0 = otherwise)
- Covariates (explanatory variables) could be included

Data description

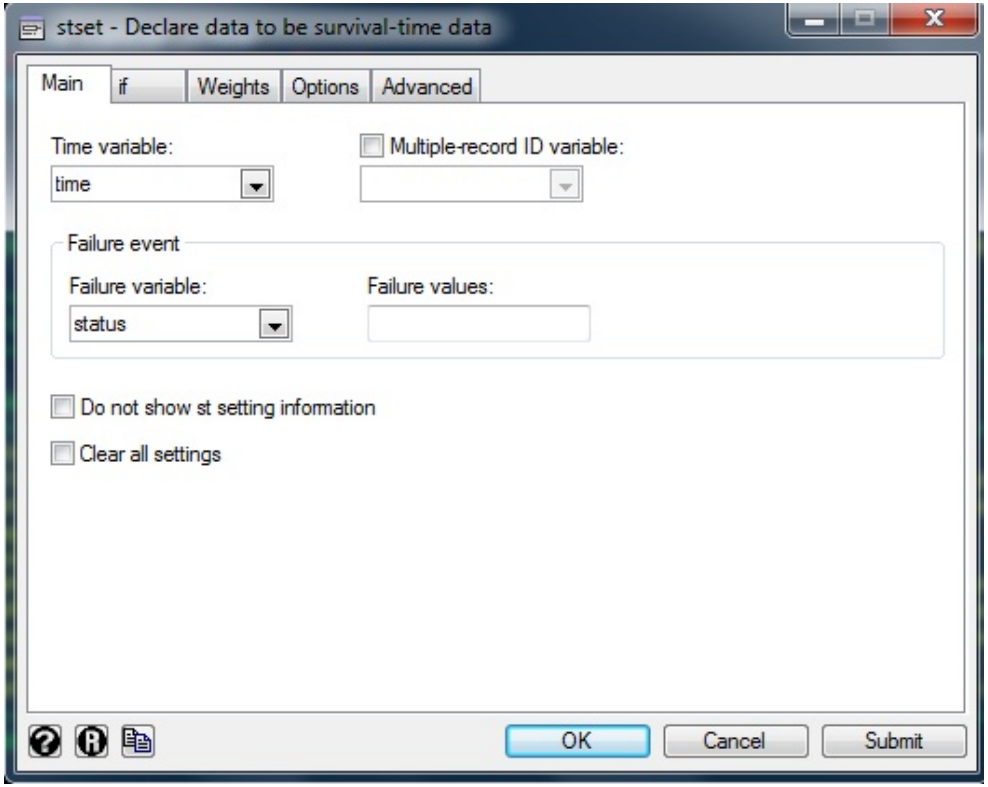
```
. use "C:\Users\user\Documents\Didattica\Southampton\SC_Epidem\lung.dta", clear

. de

Contains data from C:\Users\user\Documents\Didattica\Southampton\SC_Epidem\lung.dta
  obs:           228
 vars:            10                          28 Jan 2013 11:16
 size:         3,648 (99.9% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
inst            byte   %8.0g                  Institution code
time            int    %8.0g                  Survival time in days
status          byte   %8.0g                  censoring status 0=censored, 1=dead
age             byte   %8.0g                  Age in years
sex             byte   %8.0g                  Male=1 Female=2
phecog          byte   %8.0g                  ECOG performance score (0=good 5=dead)
phkarno         byte   %8.0g                  Karnofsky performance score (bad=0-good=100) rated by physician
patkarno        byte   %8.0g                  Karnofsky performance score as rated by patient
mealcal         int    %8.0g                  Calories consumed at meals
wtloss          byte   %8.0g                  Weight loss in last six months
-------------------------------------------------------------------------------
```

stset

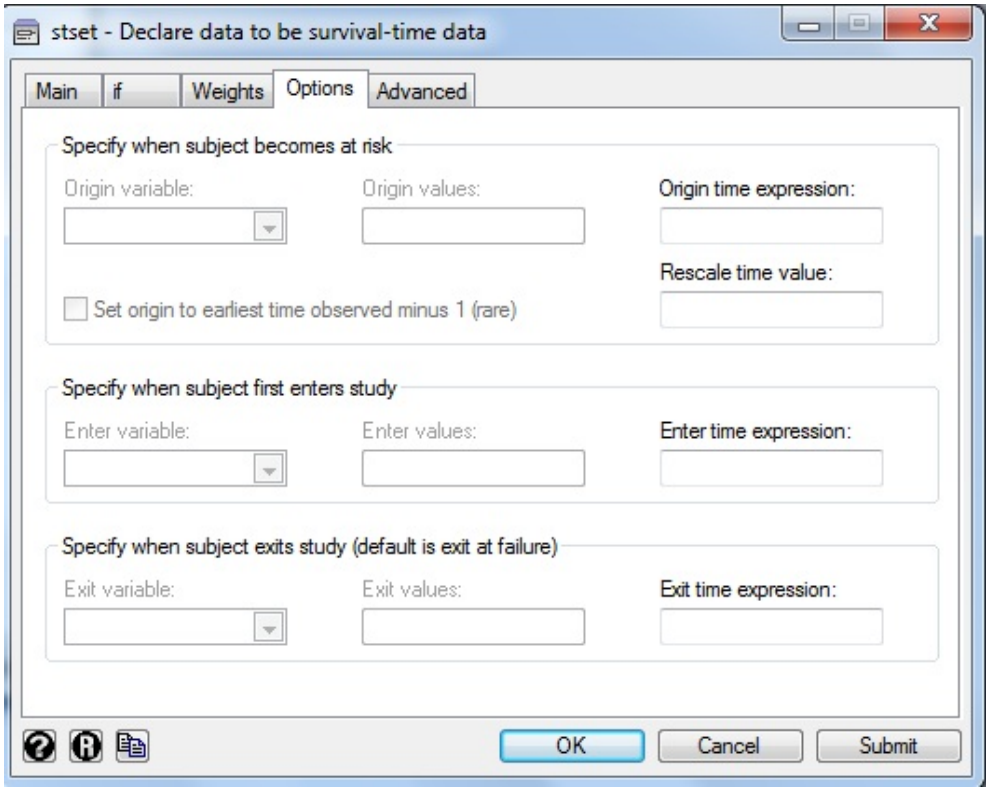
`stset` declares the data in memory to be `st` data

- Main
  - Time variable ⇒ survival time
  - Failure variable ⇒ censoring status
- Options
  - Origin time expression sets when a subject becomes at risk
  - Enter time expressions specifies when a subject first comes under observation
  - Exit time expression specifies the latest time under which the subject is both under observation and at risk.

`stset` in practice



`stset` in practice

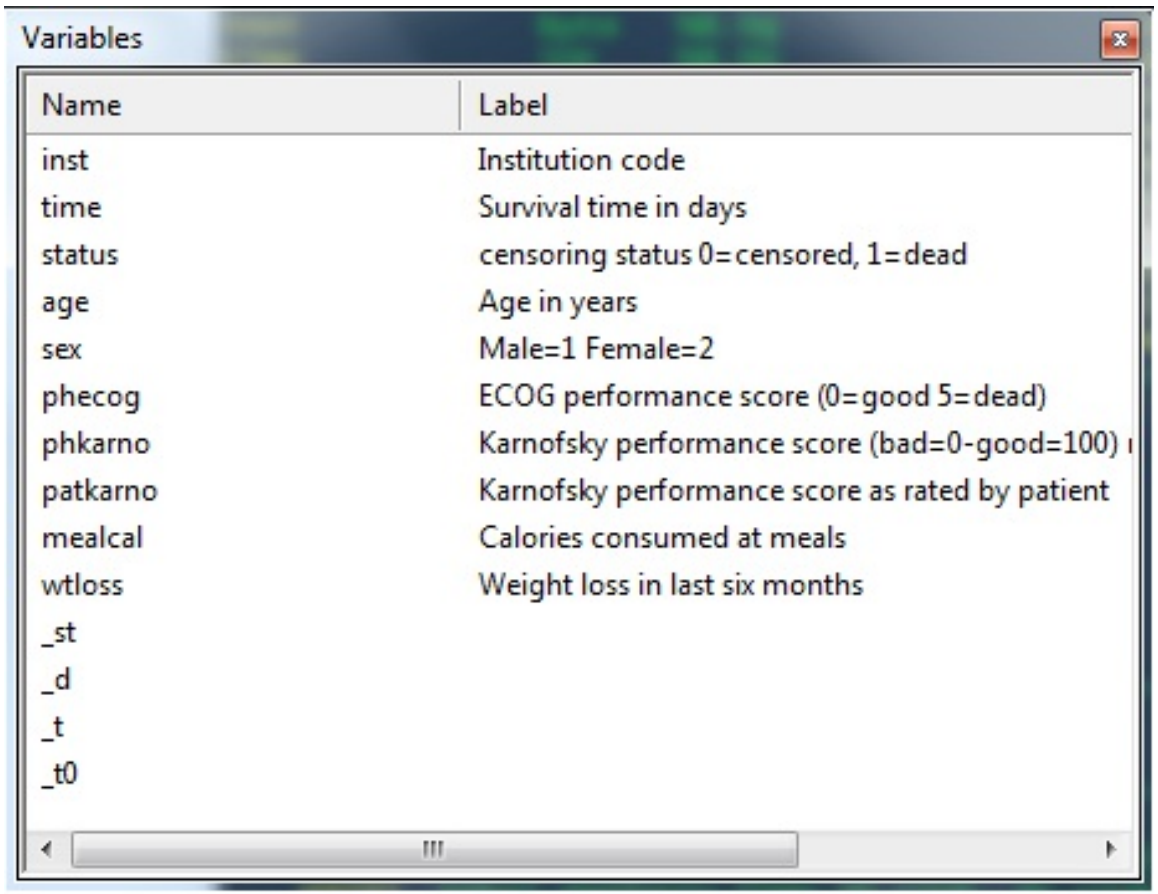


stset: example

```
. stset time, failure(status)

     failure event:  status != 0 & status < .
obs. time interval:  (0, time]
 exit on or before:  failure

------------------------------------------------------------------------------
      228  total obs.
        0  exclusions
------------------------------------------------------------------------------
      228  obs. remaining, representing
      165  failures in single record/single failure data
    69593  total analysis time at risk, at risk from t =         0
                              earliest observed entry t =         0
                                   last observed exit t =      1022
```

Using `stset`

New variables in the data, why? Which is your meaning? Should you use them?



Using `stset`

- `_st` is a binary variable indicating cases included (1) or excluded (0) from the analysis
- `_d` is a censoring indicator
- `_t` is the survival time
- `_t0` is the time at which units are observed to be at risk

Using `stset`

```
. de _*

              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
_st             byte   %8.0g
_d              byte   %8.0g
_t              int    %10.0g
_t0             byte   %10.0g

. sum _*

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          _st |        228           1           0         1          1
          _d |        228   .7236842   .4481588         0          1
          _t |        228   305.2325   210.6455         5       1022
         _t0 |        228           0           0         0          0

.
```

Summary statistics

You must `stset` your data before using

- `stdescribe` produces a summary of the st data
- `stsum` summarizes survival-time data

```
. stdescribe

        failure _d:  status
  analysis time _t:  time

                                                  per subject
Category                        total        mean         min      median         max
-------------------------------------------------------------------------------------
no. of subjects                   228
no. of records                    228           1           1           1           1

(first) entry time                              0           0           0           0
(final) exit time                         305.2325           5       255.5        1022

subjects with gap                   0
time on gap if gap                  0
time at risk                    69593    305.2325           5       255.5        1022

failures                          165   .7236842           0           1           1
-------------------------------------------------------------------------------------

. stsum

        failure _d:  status
  analysis time _t:  time

                          incidence      no. of |------ Survival time -------|
          time at risk         rate    subjects         25%         50%         75%
-------------------------------------------------------------------------------------
   total         69593    .0023709         228         170         310         550

.
```

Kaplan-Meier

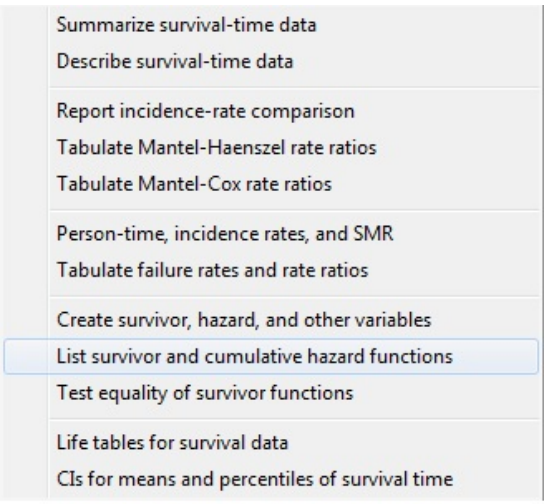
- Simple single-spell type
- Right censoring
- No left censoring (truncation)

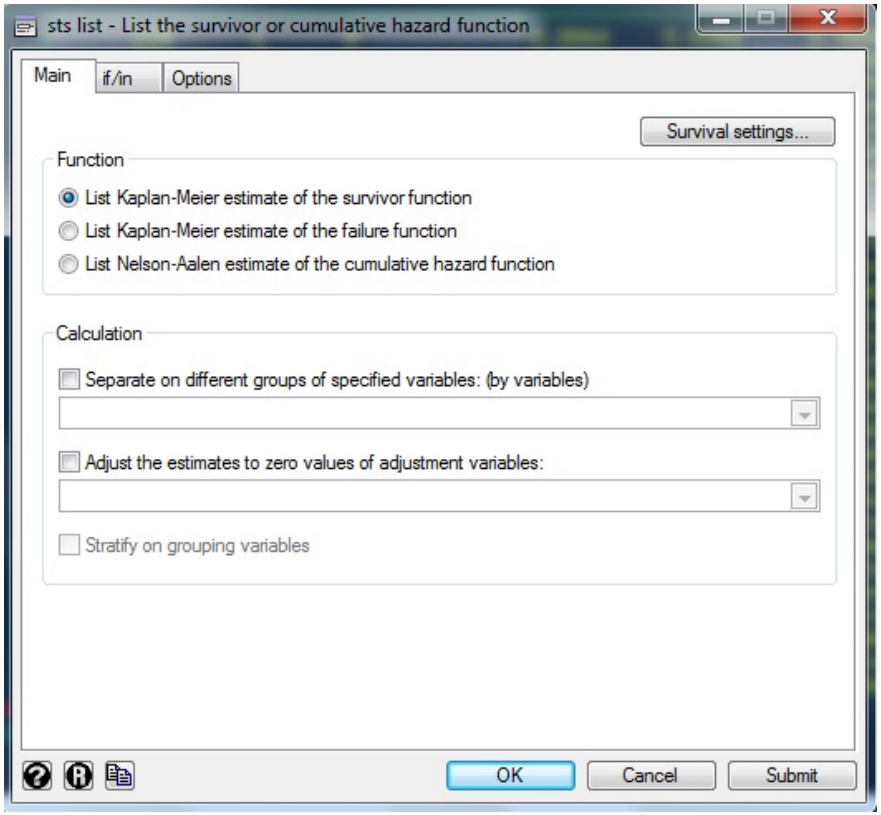
sts

Survival times are treated as observations on a continuous variable

- `sts list`
- `sts graph`
- `sts test`
- `sts generate`

sts list

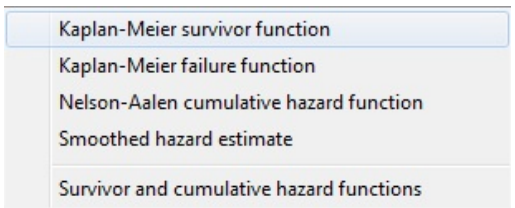

Summarize survival-time data
Describe survival-time data
Report incidence-rate comparison
Tabulate Mantel-Haenszel rate ratios
Tabulate Mantel-Cox rate ratios
Person-time, incidence rates, and SMR
Tabulate failure rates and rate ratios
Create survivor, hazard, and other variables
List survivor and cumulative hazard functions
Test equality of survivor functions
Life tables for survival data
CIs for means and percentiles of survival time

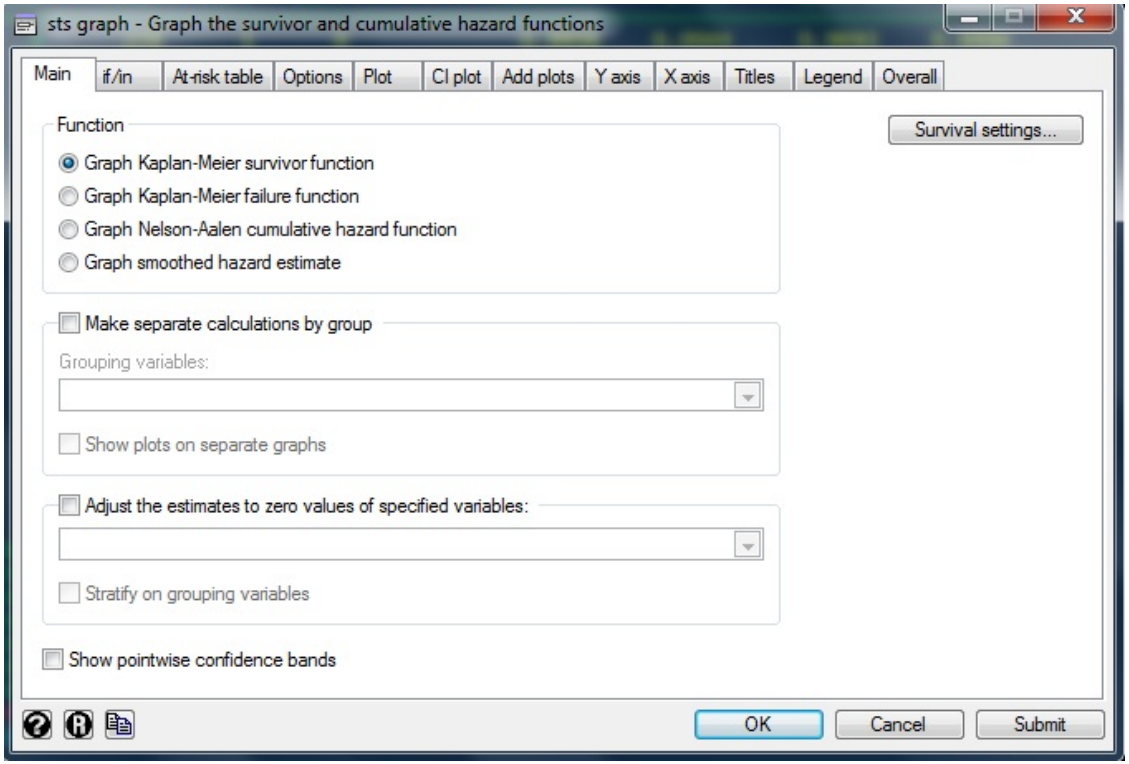

sts list - List the survivor or cumulative hazard function
Main
if/in
Options
Survival settings...
Function
List Kaplan-Meier estimate of the survivor function
List Kaplan-Meier estimate of the failure function
List Nelson-Aalen estimate of the cumulative hazard function
Calculation
Separate on different groups of specified variables: (by variables)
Adjust the estimates to zero values of adjustment variables:
Stratify on grouping variables
OK
Cancel
Submit

sts list: example

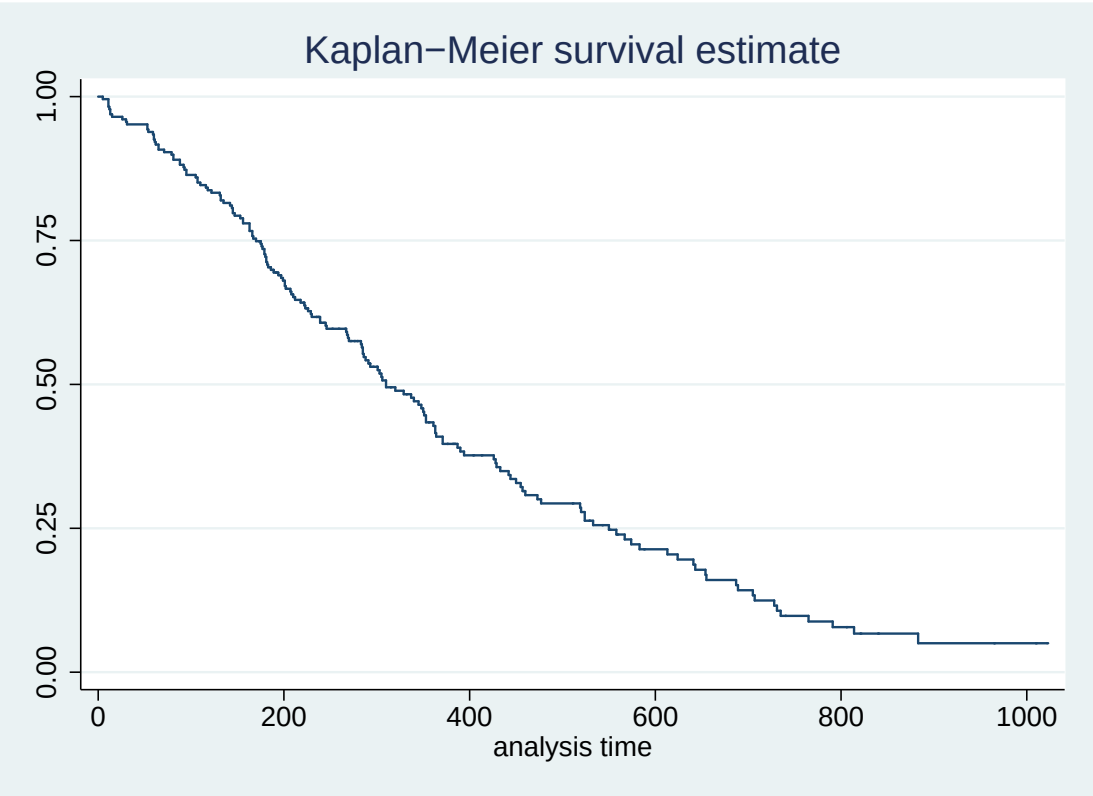
```
. sts list

        failure _d:  status
  analysis time _t:  time

          Beg.          Net      Survivor      Std.
Time     Total    Fail  Lost      Function     Error     [95% Conf. Int.]
-------------------------------------------------------------------------
   5       228       1     0        0.9956    0.0044    0.9693    0.9994
  11       227       3     0        0.9825    0.0087    0.9539    0.9934
  12       224       1     0        0.9781    0.0097    0.9481    0.9908
  13       223       2     0        0.9693    0.0114    0.9367    0.9852
  15       221       1     0        0.9649    0.0122    0.9311    0.9823
  26       220       1     0        0.9605    0.0129    0.9255    0.9793
  30       219       1     0        0.9561    0.0136    0.9200    0.9762
  31       218       1     0        0.9518    0.0142    0.9146    0.9730
  53       217       2     0        0.9430    0.0154    0.9038    0.9665
  54       215       1     0        0.9386    0.0159    0.8985    0.9632
  59       214       1     0        0.9342    0.0164    0.8932    0.9598
  60       213       2     0        0.9254    0.0174    0.8828    0.9530
  61       211       1     0        0.9211    0.0179    0.8776    0.9495
  62       210       1     0        0.9167    0.0183    0.8725    0.9460
  65       209       2     0        0.9079    0.0192    0.8622    0.9390
  71       207       1     0        0.9035    0.0196    0.8572    0.9354
  79       206       1     0        0.8991    0.0199    0.8521    0.9318
  81       205       2     0        0.8904    0.0207    0.8420    0.9245
  88       203       2     0        0.8816    0.0214    0.8321    0.9172
  92       201       1     1        0.8772    0.0217    0.8271    0.9135
  93       199       1     0        0.8728    0.0221    0.8221    0.9098
  95       198       2     0        0.8640    0.0227    0.8122    0.9023
 105       196       1     1        0.8596    0.0230    0.8073    0.8985
 107       194       2     0        0.8507    0.0236    0.7974    0.8909
 110       192       1     0        0.8463    0.0239    0.7925    0.8871
 116       191       1     0        0.8418    0.0242    0.7876    0.8833
 118       190       1     0        0.8374    0.0245    0.7827    0.8794
 122       189       1     0        0.8330    0.0247    0.7778    0.8755
 131       188       1     0        0.8285    0.0250    0.7729    0.8717
--more--
```

sts graph

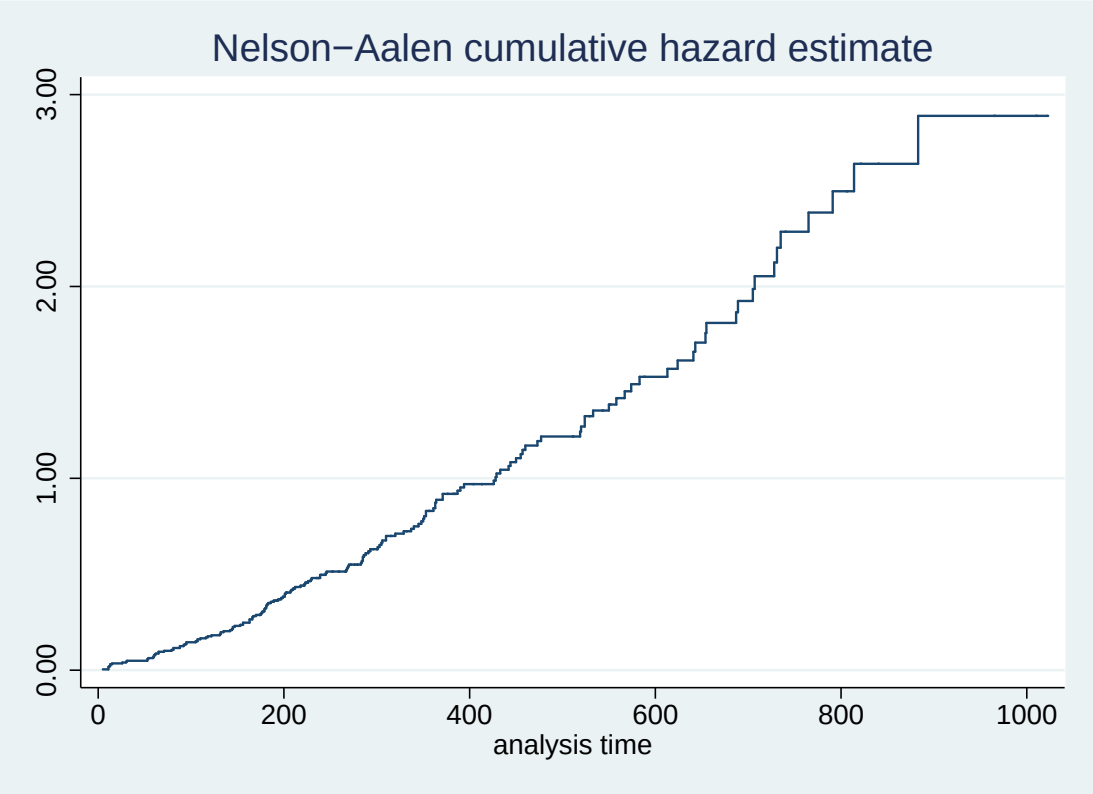

Kaplan-Meier survivor function
Kaplan-Meier failure function
Nelson-Aalen cumulative hazard function
Smoothed hazard estimate
Survivor and cumulative hazard functions


sts graph - Graph the survivor and cumulative hazard functions
Main
if/in
At-risk table
Options
Plot
CI plot
Add plots
Y axis
X axis
Titles
Legend
Overall
Function
Survival settings...
Graph Kaplan-Meier survivor function
Graph Kaplan-Meier failure function
Graph Nelson-Aalen cumulative hazard function
Graph smoothed hazard estimate
Make separate calculations by group
Grouping variables:
Show plots on separate graphs
Adjust the estimates to zero values of specified variables:
Stratify on grouping variables
Show pointwise confidence bands
OK
Cancel
Submit

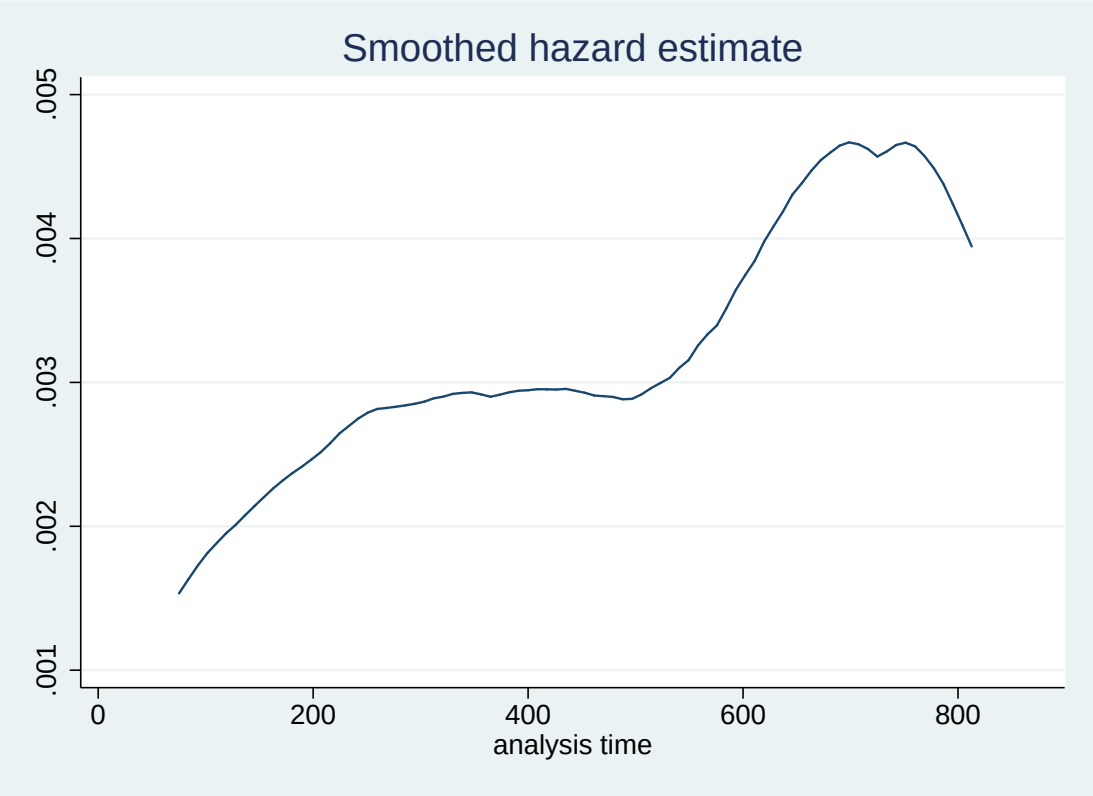
`sts graph`: example



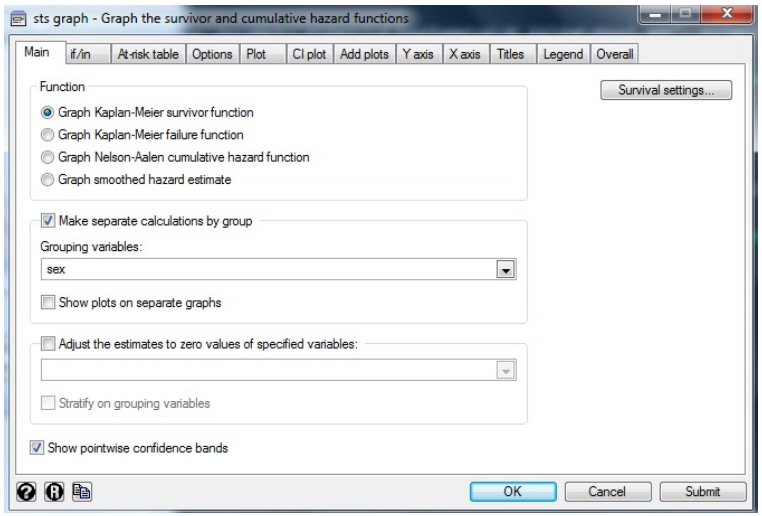
`sts graph`: example

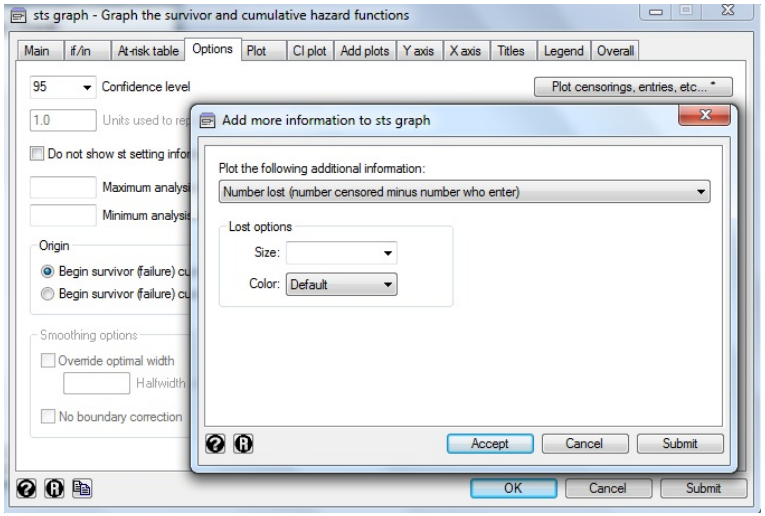


`sts graph`: example

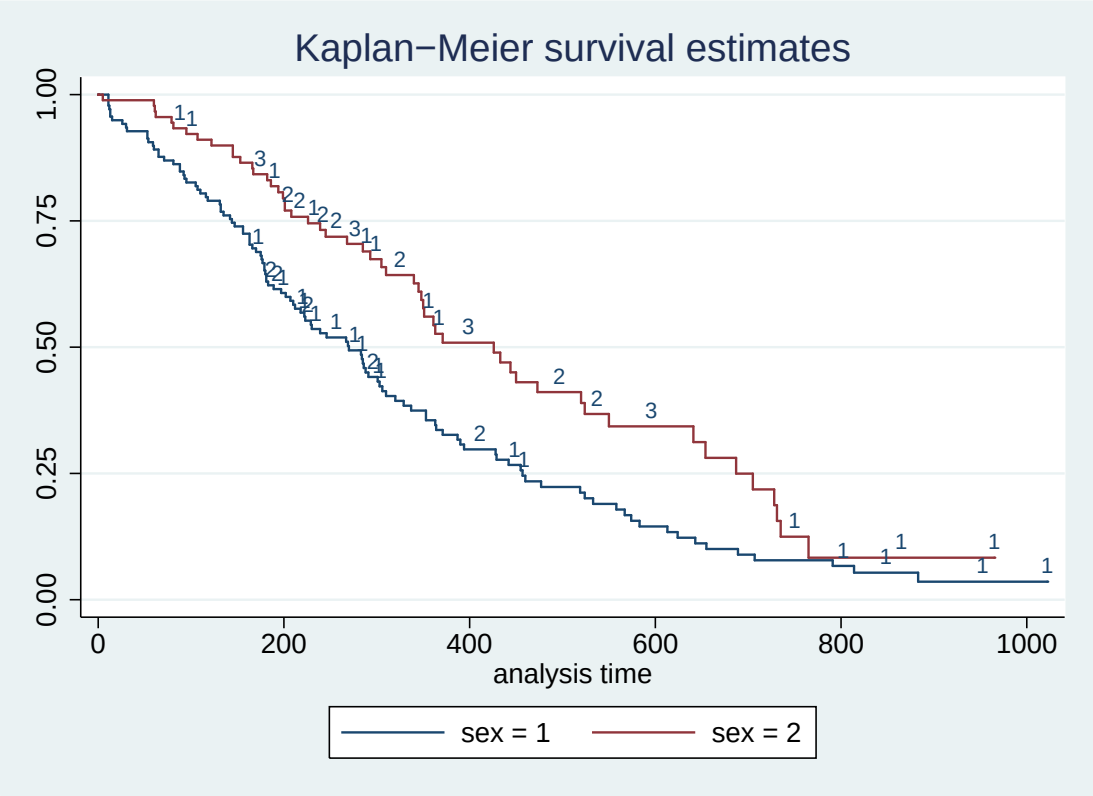


`sts graph`: stratification

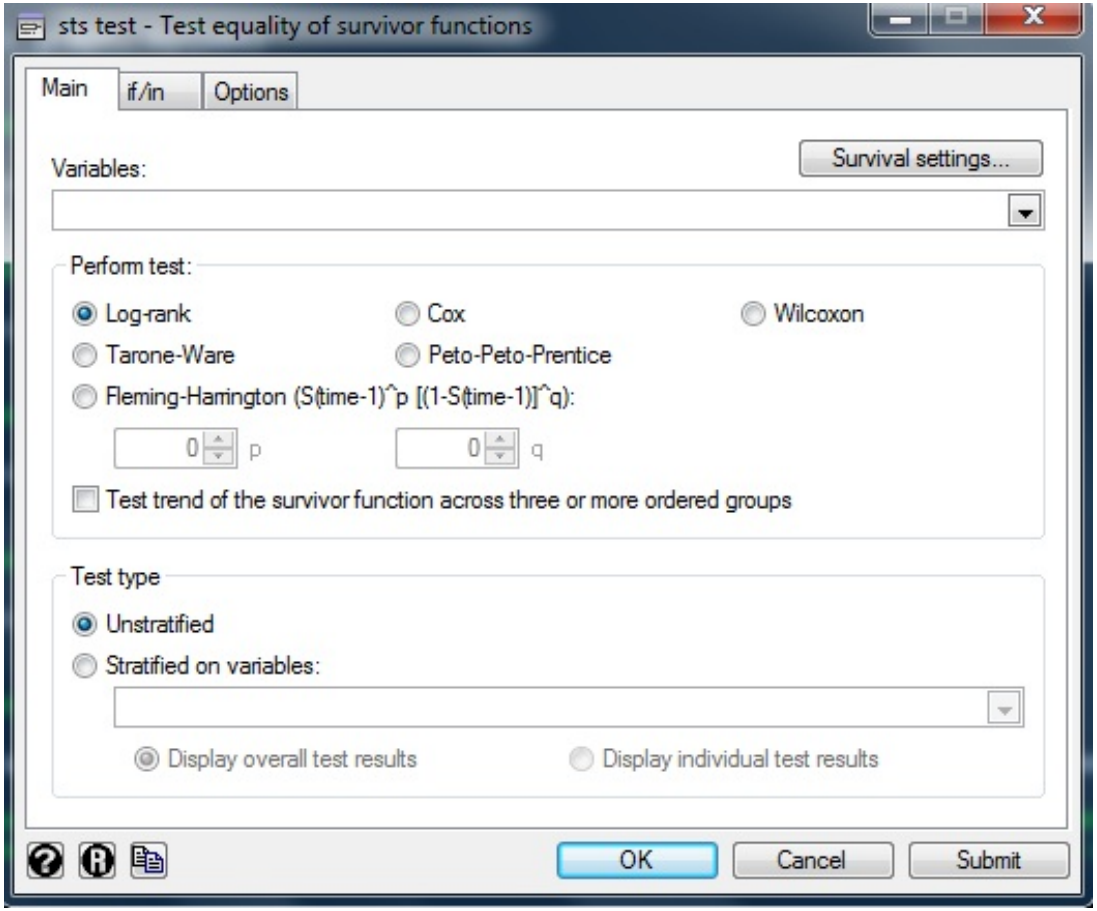




`sts graph`: **stratification**



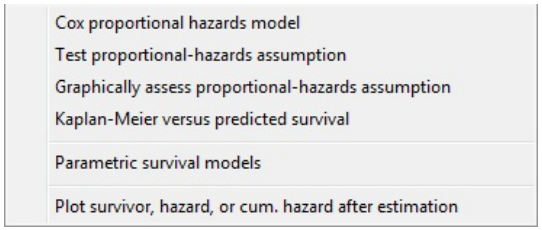
sts test

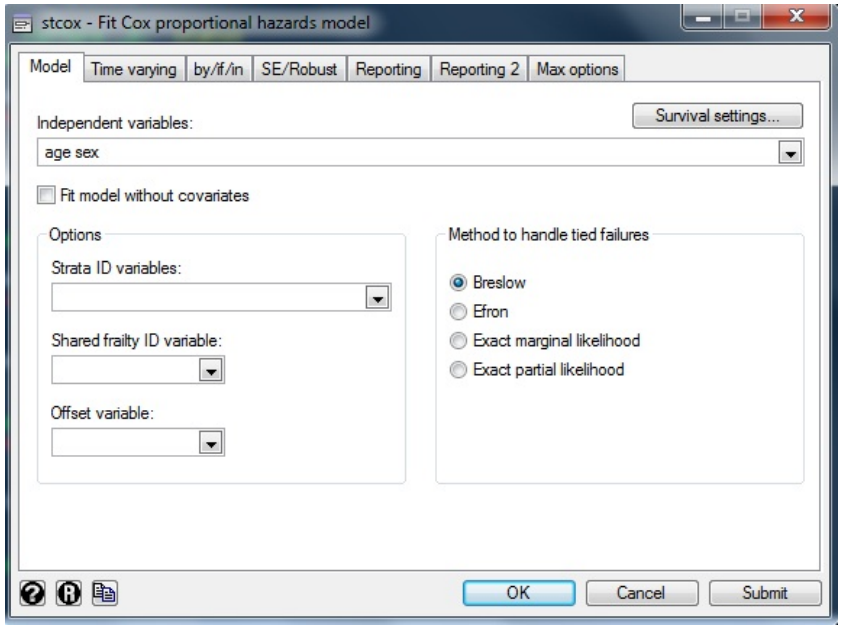

sts test - Test equality of survivor functions
Main
if/in
Options
Variables:
Survival settings...
Perform test:
Log-rank
Cox
Wilcoxon
Tarone-Ware
Peto-Peto-Prentice
Fleming-Harrington (S(time-1)^p [(1-S(time-1)]^q):
0
p
0
q
Test trend of the survivor function across three or more ordered groups
Test type
Unstratified
Stratified on variables:
Display overall test results
Display individual test results
OK
Cancel
Submit

sts test

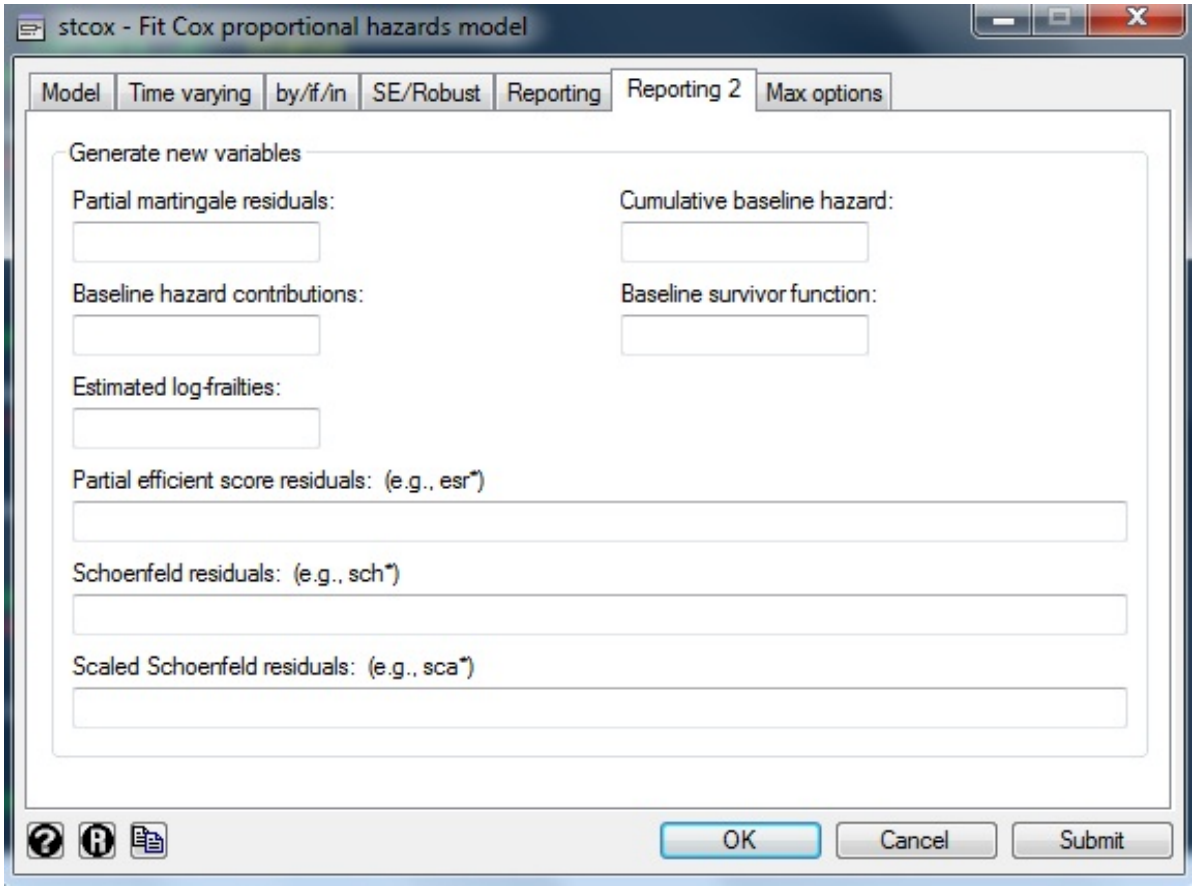
```
. sts test sex

        failure _d:  status
  analysis time _t:  time


Log-rank test for equality of survivor functions

         |  Events       Events
sex      |  observed     expected
---------+-------------------------
1        |       112        91.58
2        |        53        73.42
---------+-------------------------
Total    |       165       165.00

           chi2(1) =       10.33
           Pr>chi2 =      0.0013

. sts test sex, wilcoxon

        failure _d:  status
  analysis time _t:  time


Wilcoxon (Breslow) test for equality of survivor functions

         |  Events       Events      Sum of
sex      |  observed     expected      ranks
---------+-------------------------------
1        |       112        91.58       3148
2        |        53        73.42      -3148
---------+-------------------------------
Total    |       165       165.00          0

           chi2(1) =       12.47
           Pr>chi2 =      0.0004
```

stcox


Cox proportional hazards model
Test proportional-hazards assumption
Graphically assess proportional-hazards assumption
Kaplan-Meier versus predicted survival
Parametric survival models
Plot survivor, hazard, or cum. hazard after estimation


stcox - Fit Cox proportional hazards model
Model
Time varying
by/if/in
SE/Robust
Reporting
Reporting 2
Max options
Independent variables:
Survival settings...
age sex
Fit model without covariates
Options
Strata ID variables:
Shared frailty ID variable:
Offset variable:
Method to handle tied failures
Breslow
Efron
Exact marginal likelihood
Exact partial likelihood
OK
Cancel
Submit

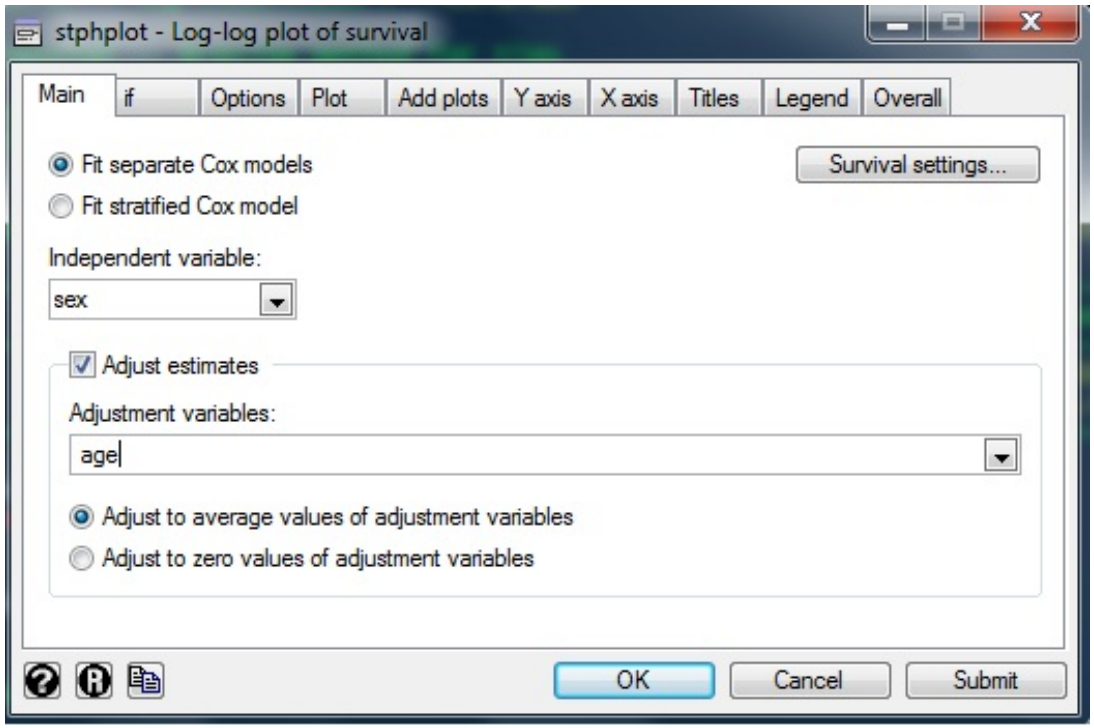
`stcox`: options for model checking



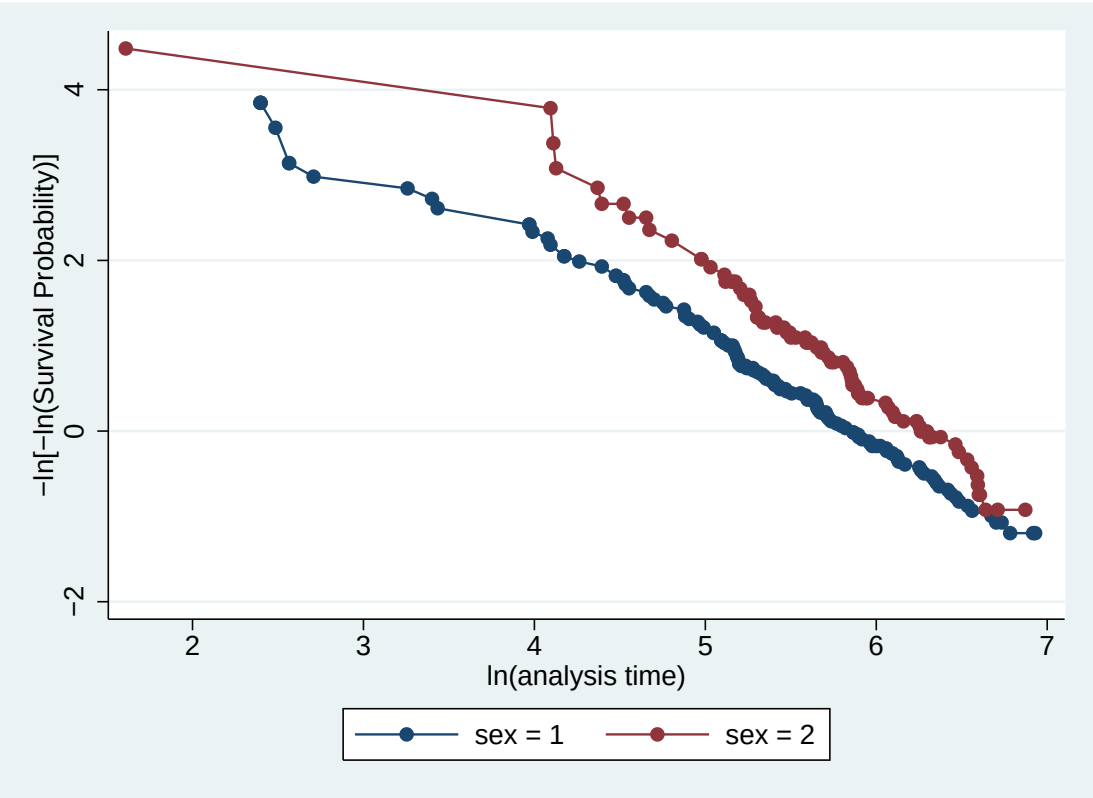
stcox: example

```
. stcox sex age,sch(global*) sca(local*)

        failure _d:  status
  analysis time _t:  time

Iteration 0:   log likelihood = -750.12202
Iteration 1:   log likelihood = -743.09465
Iteration 2:   log likelihood = -743.07965
Iteration 3:   log likelihood = -743.07965
Refining estimates:
Iteration 0:   log likelihood = -743.07965

Cox regression -- Breslow method for ties

No. of subjects =          228                 Number of obs    =       228
No. of failures =          165
Time at risk    =        69593
                                               LR chi2(2)       =     14.08
Log likelihood  =   -743.07965                 Prob > chi2      =    0.0009

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .5989574   .1003026    -3.06   0.002      .431372    .8316487
         age |   1.017158   .0093802     1.84   0.065     .9989388     1.03571
------------------------------------------------------------------------------
```

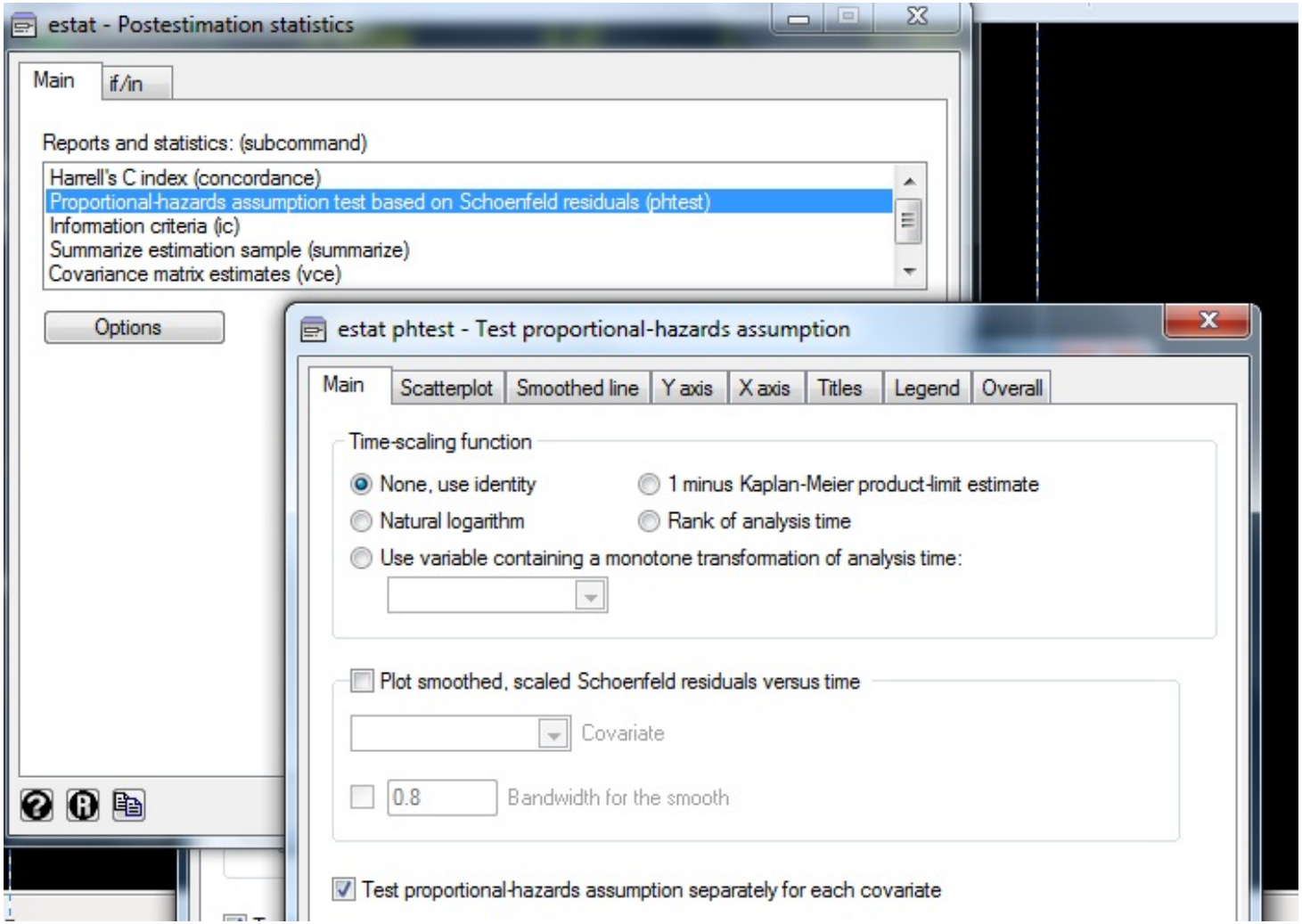
stphplot: model checking



`stphplot`: model checking



`estat phtest`: model checking



`estat phtest`: model checking

```
estat phtest, detail

    Test of proportional-hazards assumption

    Time:  Time
```

	rho	chi2	df	Prob>chi2
sex	0.12535	2.52	1	0.1125
age	-0.02090	0.07	1	0.7851
global test		2.65	2	0.2659